Final Exam: Stata Portion

EC031: Introduction to Econometrics — Spring 2026

Name: _____

Instructions:

- You may begin this portion **only after turning in the written portion** of the exam.
- Submit a single `.do` file with all your code and short answers written as comments (`*` or `//`).
- Answer the short questions **in your do-file as comments** directly after the relevant code.
- This exam is **open-note**: you may consult your own problem sets and notes, and use the Stata help menu. You may **not** use the internet, AI tools, or collaborate with others.

Load the dataset with the following command (type this as a single line in Stata):

```
import delimited "https://raw.githubusercontent.com/amichuda/ARE106data/refs/heads/master/lawsch85.csv", clear
```

This dataset contains information on **157 U.S. law schools**. The key variables are:

Variable	Description
`salary`	Median starting salary of graduates ($)
`lsat`	Median LSAT score of entering class
`gpa`	Median undergraduate GPA of entering class
`cost`	Annual tuition and fees ($)
`rank`	School ranking (1 = highest ranked)
`top10`	1 if ranked in top 10, 0 otherwise
`age`	Age of the law school (years)
`faculty`	Number of faculty members
`north`, `south`, `east`, `west`	Regional indicators

Questions

1. After loading the data, use `summarize` to find the **mean** and **median** of `salary`. What does the relationship between the mean and median tell you about the shape of the salary distribution? *(2 points)*

2. Use the `generate` command to create a new variable `log_salary = log(salary)`. Confirm it was created using `summarize log_salary`. Report the mean of `log_salary`. *(2 points)*

3. Run a **two-sample t-test** comparing mean `salary` between top-10 schools (`top10 == 1`) and all others (`top10 == 0`). Is the difference statistically significant at the 5% level? *(3 points)*

4. Run a **simple regression** of `log_salary` on `lsat`:

```
reg log_salary lsat
```

a. Interpret the coefficient on `lsat`. *(2 points)*
b. Is the coefficient statistically significant at the 5% level? How do you know? *(2 points)*

5. Now run a **multiple regression** of `log_salary` on `lsat`, `gpa`, `cost`, and `top10`:

```
reg log_salary lsat gpa cost top10
```

a. What is the R^2? What does it mean? *(2 points)*
b. Interpret the coefficient on `top10`. (Hint: the dependent variable is a log.) *(3 points)*
c. Is the coefficient on `top10` statistically significant at the 5% level? *(1 point)*

6. Compare the coefficient on `lsat` between the simple regression in (4) and the multiple regression in (5). Did it change? Briefly explain why or why not using what you know about omitted variable bias. *(3 points)*

7. You suspect that `rank` (school ranking) is a relevant omitted variable in the regression from (5).

a. In what **direction** would you expect the omitted variable bias on the `lsat` coefficient to be? Explain using the OVB formula. *(3 points)*
b. Add `rank` to the regression from (5) and re-run it. Was your prediction in (a) correct? Briefly explain. *(3 points)*

8. A classmate proposes using the school's `age` as an **instrument** for `lsat`. Evaluate this proposal by discussing:

a. **Relevance**: Is `age` likely correlated with `lsat`? Run a simple regression of `lsat` on `age` to support your answer. *(2 points)*
b. **Exclusion restriction**: Could `age` affect `log_salary` through channels *other than* `lsat`? Give one specific reason why the exclusion restriction may or may not hold. *(2 points)*