Stata Lab 4

Dataset: LAWSCH85.DTA

**Source:** Collected by Kelly Barnett, an MSU economics student, for use in a term project. The data come from two sources: *The Official Guide to U.S. Law Schools*, 1986, Law School Admission Services, and *The Gourman Report: A Ranking of Graduate and Professional Programs in American and International Universities*, 1995, Washington, D.C.

1. Use OLS to estimate: *log(salary) = β0 + β1LSAT + β2 GPA + β3log(libvol) + β4log(cost) + β5rank + β6clsize+u*

Median starting salary for new law school graduates, median LSAT score for graduating class, median college GPA for the class, number of volumes in the law school library, the annual cost of attending law school, the law school ranking with rank=1 the best, and class size.

* 1. Write your estimated model in equation format.
	2. Write out the null and the alternative hypotheses that the rank of law schools has no ceteris paribus effect on median starting salary.
	3. Use the confidence interval approach to test the null hypothesis. Calculate the 95% confidence interval (a CI such that you are 95% sure the true β falls within the CI) and confirm that it matches Stata’s output. Show all your work. Do you “reject” or “fail to reject” the null hypothesis? Is *rank* a significant predictor of salary?
	4. Now use the t-test method to test the null hypothesis. Use a 5% level of significance (95% confidence).
		1. Calculate your t-statistic and confirm that it matches Stata’s output. Show all your work.
		2. Illustrate the t-statistic relative to the t-critical value on a t-distribution.
		3. Do you “reject” or “fail to reject” the null hypothesis?
	5. Would you say it is better to attend a higher ranked law school? How much is a difference in ranking of 20 worth in terms of predicted starting salary? (Note that a ranking of 1 is the best. Also note that the dependent variable is in log.)
	6. In addition to reporting the 95% confidence interval and the t-statistic for a null hypothesis of β=0, Stata also reports the p-value for each parameter estimate. The p-value gives you the smallest α for which you could reject the null hypothesis β=0 given the t-statistic. Since a larger t-statistic makes it easier to reject the null hypothesis, a smaller p-value does the same. We prefer p<0.05, but generally p<0.10 is acceptable.
		1. What is the p-value for the estimated coefficient on *rank*?
		2. Illustrate the p-value for this problem using a t-distribution.
1. Now let’s test whether LSAT has a positive effect on salary. Use a 1-tailed t-test to test the hypothesis.
	1. Write out the null and alternate hypotheses that LSAT has a positive effect on salary.
	2. What are the t-statistic and t-critical value for a 95% level of confidence? (Note that with a 1-tailed test, the area to the right of the t-critical value is equal to α, while with a 2-tailed test, the area to the right of the t-critical value is equal to α/2. So choose the t-critical accordingly!)
	3. Do you reject or fail to reject the null hypothesis? Show your work.
	4. Does your conclusion change for a 90% level of confidence? Explain.
2. Which variables in the model are NOT individually significant in explaining salary at the 5% level of significance? Show your work using the “t-test” or “confidence interval” method.
3. Sometimes several variables are individually statistically insignificant, but they may still have joint significance. Before removing them from our model, we need to test for this joint significance, i.e. do the variables together explain some variation in the dependent variable?

Are the individually insignificant variables in #3 jointly significant in explaining salary? To test joint significance you need to use an F-test that compares the added explanatory value of the unrestricted model to the restricted model.

Unrestricted model: the regression in #1

Restricted model: the regression assuming your null hypothesis is true

F-statistic= [(R2UR-R2R)/m]/[(1-R2UR)/(n-k)], where m=# of restrictions, (n-k)=degrees of freedom for unrestricted model.

Note that the F-statistic will always be positive, since the R2 never decreases when there are *more* explanatory variables in the model. We are testing whether the R2UR is sufficiently larger than the R2R to warrant keeping the individually insignificant variables in the model.

1. Write out the null and alternate hypotheses.
2. Write out the restricted model.
3. What is the R2 from the unrestricted model? From the restricted model?

(When estimating the restricted model, be sure to account for missing data on *LSAT* and *lcost*. You want to use the same observations in your unrestricted and restricted model. Use the expression:

 *reg y x’s if LSAT!=.&lcost!=.*

where != means “not equal to”, the . indicates a missing value for a cell in the dataset, and & indicates “and”. The included x’s are only those from the restricted model.)

1. What is the F-statistic? What is the F-critical value for 5% level of significance (for 95% confidence)?
2. Illustrate the F-statistic relative to the F-critical value on an F-distribution.
3. Do you reject or fail to reject the null hypothesis?
4. What is your conclusion regarding keeping the individually insignificant variables in the regression?
5. Let’s remove *LSAT* and *clsize* from the original model in #1.
	1. What happens to the statistical significance of *llibvol* if you remove *LSAT* and *clsize* from the original model? Explain why this happens. (Consider the formula for the t-statistic.)
	2. Using your new model, state and test the null hypothesis that *log(library volume)* and *log(cost)* have equal effects in predicting starting salary, i.e. that a 1% increase in library volume has the same percentage effect on salary as a 1% increase in cost. (To generate a new variable in Stata, use the command *gen* by typing: *gen varname=function)*
		1. Null and alternate hypotheses:
		2. Unrestricted model:
		3. Restricted model:
		4. Test statistic, critical value, conclusion: