ECON 453 - Econometrics
Fall 2023
Exam 2 Practice Problems

ANSWER KEY

1. A – A linear probability model is an OLS model in which the dependent variable is binary. We are intending to estimate the probability that an individual is a value of 1 ("Yes") for the dependent variable.

2. B

3. D – an instrumental variable should directly influence the explanatory variable of interest and *not* directly affect the dependent variable.

4. B.

5. C – This was part of our discussion of panel data/fixed effects analysis. The idea is that cross-sectional variation on these types of relationships can provide misleading results due to omitted variable bias.

6. D. The y-variable in this model is now specified as the *change* in the approval rating from month to month. The explanatory variable is the change in the unemployment rate between months. This means the constant (intercept) tells us that we predict a slight drop in the approval rating (0.8 percentage points) each month if the unemployment rate did not change from the previous month.

7. B – The first step is to estimate the change in the approval rating. This is estimated as -0.8-(2.6*change in unemployment). The change in unemployment from October 2016 to November 2016 is -1 (dropped from 5 to 4). This means our predicted change = -0.8 – (2.6*-1) = -0.8+2.6 = 1.8. Our model predicts the approval rating will increase by 1.8 percentage points in November 2016. Since it was 45 in October 2016, this means we predict it will be 46.8 in November.

8. A

9. A

10. Estimating who has children (linear probability model)

    a. The coefficient on age tells us that for every year older a person in our sample is, all else equal, we predict the likelihood he or she has kids to increase by 4.3 percentage points. Females in this age group are, all else equal, 20.8 percentage points more likely to have their own kids that live with them, and those with a bachelor's degree are about 12 percentage points less likely to have children at this age than those without a degree. One thing that is important to keep in mind here is that we are looking at a specific age range (25 to 35). This is part of why the coefficient on bachelor's degree is so large. People with a college degree are more likely to have kids at older ages, so this coefficient would likely be smaller in magnitude if we looked at a sample of 25 to 45-year-olds, for example. The female coefficient is quite large. There are at least a couple of factors that explain this. First, as we discussed, in the average male-female couple, the male tends to be a little older than the female. Meaning some of the females in our sample will be married/ dating/chillaxing with partners that are too old to be in our sample. Secondly, the way the Census measures this variable is based on whether or not the children live in the same house as you. In the U.S., it is far more common to observe a single-parent household headed by a female than headed by a male.

b. To predict y-hat, we can plug into the equation with age=33, female=0, and bach=1:

$$\hat{y} = -0.923 + 0.043(33) + 0.208(0) - 0.120(1) = 0.376$$

This means that our model predicts that there is a 37.6% chance this person has children (that live in the same household as he does).

c. The interaction term tells us that the effects of having a bachelor's degree on whether or not a person has children is quite different for males and females. In particular, among 25 to 35-year-olds, having a college degree decreases the probability a male has children by a little bit, but it decreases the probability a female has children by quite a bit. To see this more clearly, we can plug in for the female dummy (0 for males, 1 for females) and create separate equations.

For males: $\hat{y} - .944 + .043age + .260(0) - .042bach - .150(0 * bach)$

$\hat{y} - .944 + .043age - .042bach$

For females: $\hat{y} - .944 + .043age + .260(1) - .042bach - .150(1 * bach)$

$\hat{y} - .684 + .043age - .192bach$

So, for males, having a bachelor's degree is estimated to reduce the likelihood of having children in your home by 4.2 percentage points, all else equal. For females, the number is much larger in magnitude, with an estimated impact of the bachelor's degree of 19.2 percentage points.

d. Overall, in our sample 43.9% (145,748/331,745) have children. We predicted that 36.9% did (122,260/331,745). Among the 185,997 people that actually did not have children, we predicted correctly for 140,417 of them. This represents an accuracy rate of 75.5%. Before you get too excited about that, let's check how accurate we were at predicting for the 145,748 people that do actually have children. We predicted correctly 76,680 of these, for an accuracy rate of 52.6%. Overall, we were correct in predicting 65.4% of our sample ((140,417+76,680)/331,745).

11. Estimating who has children in Idaho (Logit model)

a. One thing I am looking for here is that you know the difference between the "raw" coefficients that come from estimating a logit model, and the "marginal effects" that we use to give more meaningful interpretations. In gretl results, the part we want to use for practical interpretations are the numbers in the "slope" column. These tell us the impact on the probability of having children for a one-unit change in each variable, holding all else constant at average levels. The marginal effects estimates tell us that, for the average person, each additional year of age increases the likelihood of having children in the home by 4.5 percentage points. This seems like a large number, but the ages of 25 to 35 are where a lot of the "kid-action" is. This means that we do see big changes during this age range. The income variable is positive and significant, but very small in magnitude. The female variable is large, positive, and significant. This tells us that females, all else equal, are expected to have a probability of having children in their home of about 20 percentage points higher than males. The likely explanations for this were discussed in the previous problem. Finally, we find out that having a bachelor's degree means it is less likely that a person in this age range has children (by about 7 percentage points). This might be because people that go to college delay starting a family. Another explanation might be that people that go to college learn about what it actually takes to produce a baby and just decide it is not a good idea.

b. We can get a predicted probability by plugging the intimate details of T-bone's life into the logit "nasty equation". Note that you need to use the raw coefficients in this estimate:

$$\hat{y} = \frac{e^{-5.640+(.185*32)+(.004*45)+(.851*0)-(.283*1)}}{1+e^{-5.640+(.185*32)+(.004*45)+(.851*0)-(.283*1)}} = 0.544$$

Our equation predicts there is a 54.4% chance that T-bone lives with at least one child in his home. To find what happens to our prediction if he did not get his college degree, we just change the equation to reflect the difference:

$$\hat{y} = \frac{e^{-5.640+(.185*32)+(.004*45)+(.851*0)-(.283*0)}}{1+e^{-5.640+(.185*32)+(.004*45)+(.851*0)-(.283*0)}} = 0.613$$

So, we see that if he did not have his college degree, the likelihood that he lives with children is now estimated at 61.3%. This means the college degree decreases the likelihood of living kids by 6.9 percentage points (61.3-54.4). Note that this is very similar to the estimated effect of the college degree in the marginal effects table. It is not always the case that these estimates will match since those in the table are the average results and the one we calculated by hand is for a specific individual (T-bone).

12. Consider a 2017 study in the Journal of the American Medical Association that examines whether legalization of recreational marijuana in Colorado and Washington affected adolescent usage.
    a. The difference-in-difference estimates here should be comparing each state (Colorado, then Washington) to the other states. So, for Colorado, the difference is 0 (rate did not change before and after), and for Washington, the rate went up by 2.0 percentage points (8.2-6.2). In the control states, the usage rates decreased by 1.3 percentage points. This says then that the difference-in-difference estimate for Colorado is usage increasing by 1.3 percentage points, and in Washington the effect of legalization was an increase in usage by 3.3 percentage points. To think about it a slightly different way: the control states tell us usage was generally dropping among 8th graders over time. We expect this 1.3 percentage point decrease is what would have happened in Washington if not for the legalization. Instead, Washington experienced a 2.0 percentage point increase, so this was 3.3 percentage points more than expected (2.0 – (-1.3)).

    b. The key thing for difference-in-difference analysis results to be believable is that we consider the control group to be the appropriate counterfactual to the treatment group. Do you believe that using the average usage rates in all other states is a good basis for setting the expectations of what would have happened in Colorado/Washington absent legalization? One thing about selecting all states is it takes out some of the "cherry picking" that can occur if we try to pick and choose which states to include as controls. You could argue that we should pick a more similar, smaller set of states for the control group. If you think that, then you would need to consider how to choose the set of control states. Should we look at the other states in the West region? Should we look at states that had the most similar usage rates in the before period?

    c. One reason we might try to look at earlier data is to try and investigate the "parallel paths" assumption of the difference-in-difference technique. This could help us, for example, with choosing the control group discussed in part b of the question. Are the trends in usage among 8th graders on a similar trajectory over time heading into the policy change? If yes, that helps convince us that the comparison is valid, and we are identifying the actual impacts of the policy change on usage rates.

13. Time-Series: U.S. Life Expectancy

a. The first thing we notice is that our adjusted R^2 and t-stat values seem to indicate that we are some sort of statistical gods/goddesses. Then we remember that this is a time-series model and we have probably not really found a way to explain 99.3% of the variation in life expectancy in the U.S. over time. Each of the coefficients is highly significant, but some of the results are confusing. For example, this indicates that increases in GDP per capita by $1,000 (all else equal) will decrease life expectancy by 0.11 years. Furthermore, an increase in cigarettes smoked per person by 1,000 will increase the average life expectancy by 0.58 years. The coefficient on the bachelor's variable says that if an additional percentage point of the population has a college degree, life expectancy increases by 0.62 years. Overall, these do not really match my expectations. Based on this model, I suggest we all start smoking a lot more cigarettes than we currently do. Don't argue with me on this one unless you can explain why I got such a strong p-value indicating a positive effect. Oh, what's that? You say this is a classic case of a spurious result and lack of accounting for trends? Ok then, you sound like you know what you are talking about. The validity of this model leaves much to be desired.

b. The null hypothesis for these tests is that there is no serial correlation in our error terms. The p-value in the test for positive autocorrelation (serial correlation) is very low (0.0071632), which tells us we must reject the null. We have a severe problem that needs to be dealt with before we can reach any conclusions about these relationships.

c. When we include the time trend variable, the results do not really seem to improve or even change in any meaningful way. The trend variable is not significant, and the coefficients on the other variables are all pretty similar to what they were in part a. We still have confusing results for both the GDP per capita and cigarette consumption variables. Sometimes including a time trend variable will improve the results of our models, but this does not seem to be the case in this particular model.

d. Including the lagged value of the dependent variable is another way to try to account for underlying trends in time-series analysis. In these models, we control for the previous period's level of Y and try to explain whether the x-variables explain the remaining (unexplained) variance in current period Y. I also decided that we were likely to have a serious issue with multicollinearity, so I split into three separate models. In these models, we see that the percentage of the population with a bachelor's degree is the only variable that has any explanatory power, after accounting for the lagged value of life expectancy. The coefficient on the number of cigarettes consumed is a positive value but is not at all statistically significant.

e. We discussed this issue specifically a while ago in class. The problem in trying to find what "explains" changes in life expectancy is that life expectancy grows at a pretty constant rate in most advanced economies (*before recent years). The easiest thing to use to predict how life expectancy will change next year is to predict it will go up a little bit. Since there is very little deviation from this trend, there is very little independent variation for us to identify the impact of explanatory variables such as cigarette consumption. There are many ways that one could approach the general question of the relationship between cigarette consumption and health differently. To begin with, we might want to use a different health outcome measure than life expectancy. We could use rates of certain types of cancer, for example. Another thing we could probably do is use some cross-sectional variation rather than relying on time-series. Ideally, we would get data about individual cigarette use and health outcomes, but that might be hard to get. Perhaps instead we could look at differences in rates of smoking across cities and states and compare how these relate to certain health outcomes. We could also utilize panel data if we are able to collect information for states over time. Another route would be to compare across

countries (and again, perhaps we could use panel data here). This is the type of question where there is not one specific answer that I am looking for. What I want to see is that you are thinking about the "big picture" issues we have in our models and how we might adjust things to get some estimates that do a better job answering the question we have in mind.

14. University Panel Data!

a. The coefficient on the tuition variable indicates that if tuition increases by $1,000, all else equal, the enrollment is expected to increase by about 245 students (.245 of 1,000). This is significant at the standard 5% level of significance. However, this does not at all match my expectations. If I know anything about economics, it is that a higher price is supposed to decrease the quantity demanded. The coefficient on the "public" dummy variable indicates that public universities tend to have larger enrollments than private universities. The estimate tells us the typical difference is more than 18,000 students. This seems a little high, but I would definitely expect this coefficient to be positive as there are some very large public universities around.

b. The use of panel data means we have a lot more observations to work with. However, when we run the pooled OLS model, we get very similar results. Tuition is still estimated to have a positive and significant impact on enrollment, which is distressing. The estimate of the difference between public and private schools is very similar as well. Overall, not much has changed when we ran a pooled model using our panel data.

c. In part c, we make use of the panel data and use a "fixed effects" specification. As discussed in class, you can think of this as running a model that includes dummy variables for each particular year and each particular institution. The main reason to do this is to control for omitted variables that were causing a very significant bias in the results in part a. The problem in our earlier models is that there are many ways in which universities differ, and we were not controlling for any of them other than tuition and public/private status. The best schools tend to have large enrollments and high tuition. If we don't account for the differences in quality, region, and other factors, then it makes it look like the tuition being higher causes the enrollment to go up. In fixed effects estimation, we are now focusing on how tuition changing within an institution impacts enrollment at that institution. So, for example, we see how enrollment at UI changes as tuition at UI changes over the 9-year period of our sample. We are holding all time-invariant school-level characteristics (like region, public/private status, etc.) constant. When we do this, the results make a lot more sense. We now get a negative and significant coefficient for our tuition variable. This tells us that if tuition increases at the typical school by $1,000, we will see the enrollment at that school decrease by about 147 students. This result is much more like what I would have expected, and my faith in the law of demand (and Economics as a discipline) has been restored.

d. The dummy variables for each year are included to pick up the general trends in our y-variable throughout the period of our panel dataset. We interpret each of the year dummies relative to the omitted year, which is 2003. This tells us, for example, that we can expect the typical research university to have about 123 more students in 2004 than in 2003, holding all else equal. By 2011 (year 9 in the sample) the average enrollment in these institutions is expected to be about 2,021 students higher than in 2003. Overall, the dummies tell us there has been a statistically significant increase in the average number of students enrolled in this set of universities over time.