ECON 453 - Econometrics
Fall 2023
Exam 1 Practice Problems

ANSWER KEY

1. **C** – This would be a test for the equality of coefficients. The null in these tests is that the coefficients are equivalent to each other, we are looking to see if we have evidence that they are not.

2. **D** – do you see any with a log transformation? No? then how dare you try and say "percent". A 1 unit change in the share that has a college degree would be a 1 "percentage point" change (from 30 percent to 31 percent, for example). The other part is recognizing that the y-variable is scaled in 1000s of $.

3. **A** – We have an endogeneity problem when are variables are correlated with the residuals.

4. **D** – Unbiased means the expected value of our estimate (with the hat) is the true value in the population (beta without the hat).

5. **B.**

6. **A**

7. **C** – females are the "reference category" in this one. The coefficient on the interaction term will tell us how to adjust impact of having a bachelor's degree for males.

8. **C** – As the previous answer mentions, the coefficient on the interaction is the one telling us how to adjust the impact of bachelor's degrees for males. If this coefficient is statistically significant, it tells us the impact of a degree is different for males than for females.

9. **False**. There is no way to "correct" standard errors for a multicollinearity problem. This is one of our options when dealing with heteroskedasticity.

10. Suppose we are interested in estimating the relationship between the annual income in a household and the annual expenditures on clothing for a household. We collect a dataset that includes this information for a random sample of 1,000 households in 2022. Our explanatory variable, annual household income, is reported in 1,000s of US$. Our dependent variable, annual household spending on clothing, is also measured in 1000s of US$.

    a. We don't really need to worry about the interpretation of the intercept here, particularly because it doesn't make much sense. It tells us that if a family has $0 in annual income, we expect them to spend -$500 on clothing. I want you to focus on the slope coefficient. Technically, this says that if x increases by 1 unit, we expect, all else equal, that the y variable will increase by .025 units. To put this in practical terms, we would say that if household income increases by $1,000, then, all else equal, that household is expected to increase its clothing expenditures by $25. This seems like a pretty reasonable prediction to me, but then again, I live in a household that does not place a high priority on clothing expenditures.

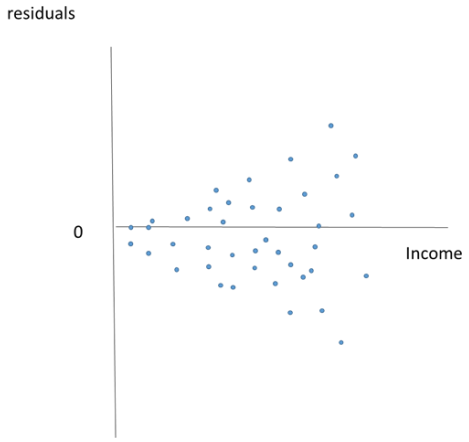
    b. To predict the clothing expenditures for the Davidson family, we can simply plug into our equation from part b. Remember that the units here are thousands of dollars (for both the x and the y variables).

$$\hat{y}_i = -0.5 + 0.025x_i$$
$$\hat{y}_i = -0.5 + 0.025(110) = 2.25$$

This means our model predicts the Davidsons spent $2,250 on clothing in 2022. Since we know they actually spent $5,000, our residual is $2,750 (or 2.75 in the units of the equation). This means the Davidsons spent quite a bit more on clothing than we would expect based on their income. Perhaps some of this spending is being done to cover up the emotional scars from little Donnie's accident, but that is probably a discussion for another day.

c. The idea here is to think about building a model around the y-variable. We want to predict how much a household will spend on clothing, and we only have one explanatory variable at this point: how much income the household has. To me, the most obvious variable to add would be the number of people in each household. I would expect, all else equal, households with more individuals will have a higher expenditure on clothing because they have more people to cover. Another thing I think we might want to add would be something more about the demographics of the household. We would probably want something, for instance, that gives us an idea of the ages of the people that live in the household. We could start with something like number of kids, number of people that are over 65, etc. There are lots of variables you could include here as long as you think that: (a) they are influential in how much a household will spend on clothing, and (b) they will be different for different households in the same year. As a few more examples you could include the rural/urban setting of the household (I would think households in urban environment spend more on clothes), the climate in which the household resides, or the number of relatives the members of the household have that work for beer distributors or radio stations.

d. The relationship between the explanatory variable of income and the response variable of clothing expenditures is one example we used to discuss the assumption of Homoskedastic errors. This assumption means that we want to see constant variance in our errors as the values of each x variable (in this case, income) change. If we do not have this, we have heteroskedastic errors, which means that our statistical inference information, such as t-statistics and p-values, might not be reliable. The reason we expect the problem of heteroskedasticity in this example is because we expect those with relatively low incomes to have relatively low variance in clothing expenditures. Since they do not have much income, most households at this level spend similar amounts on clothing. When we look at higher income households, however, there is a lot more variation. Some high-income households will choose to spend a lot on clothing, while others might spend very little and choose to spend on other types of things. In terms of the residuals, this means that we will likely not be very far off (in either direction) in our predictions for low-income households, but that our predictions for high-income households could be very far off in either direction. An example of this problem would look something like:



11. Consider the dataset of males between the ages of 28 and 35 that contains information on weekly earnings and a number of other characteristics. We have worked with this dataset in class. Suppose you want to develop a model of the following form:

$$y_i = \beta_0 + \beta_1 educ_i + \beta_2 \exp er_i + \beta_3 age_i + \beta_4 IQ_i + \mu_i$$

a. We can use the estimated coefficients from the table above and write as the following equation:

$$\hat{y} = -757.61 + 42.97 educ + 12.46 \exp er + 11.93 age + 5.84 IQ$$

To give numeric interpretations of the coefficients, we can say:
- All else equal, a 1-year increase in education is expected to raise weekly wages by $42.97
- All else equal, a 1-year increase in work experience is expected to raise weekly wages by $12.46
- All else equal, a 1-year increase in age is expected to raise weekly wages by $11.93
- All else equal, a 1-point increase in IQ is expected to raise weekly wages by $5.84.

It certainly makes sense to me that each of the variables in the model would have a positive relationship with wages. I find it a little surprising that there is such a large difference between the values of a year of education and a year of work experience. We should note that the education, experience, and IQ coefficients are all highly statistically significant, but the age variable is not (p-value of around 0.12). This makes sense to me as well, once we have accounted separately for years of work experience, I am not really sure why age would affect wages one way or another. This is especially true given the relatively narrow age range of this sample (28-35).

b. The F-statistic that is provided in standard OLS results is the test for the overall joint significance of all of the variables in the model. This is a test of the following hypotheses:

$$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$$
$$H_A : \text{variables are jointly significant}$$

The p-value in this case is approximately 0.000000000000000000000000482, which I take as fairly strong evidence that we can reject the null. This means our model has overall explanatory power. As we discussed, this is not a very stringent test. It is essentially saying that our regression model is better than simply guessing the average wage for each individual when trying to predict wages.

c. Just from looking at the variables included in this model (before seeing the correlations), I would be somewhat concerned about multicollinearity. For example, I would expect education and IQ to have a strong positive correlation. I would also expect the age and experience variables to be fairly highly correlated. When I look at the actual results of the correlation table, however, it doesn't seem as big of a problem as I initially suspected. The education and IQ variables are fairly strongly correlated, but an r of 0.489 is not an immediate cause for concern. The next biggest correlation is the negative one between education and experience, but again, nothing too worrisome. Overall, the correlation table does not make me think we need to be overly concerned with multicollinearity. In order to further consider this issue, we could calculate the variance inflation factors (VIFs) for each of the variables in the model. These tell us how well we can predict a given x variable based on the other x variables. For example, if we know a person's experience, IQ, and age, how well can we predict that person's education level? If we can predict the education very well, then we will get a high VIF value (above 5 (somewhat concerning) or 10 (very concerning) and are likely to have a substantial multicollinearity problem.

d. Generally, we are thinking about omitted variable bias in the case where we cannot collect data for a certain (meaningful) variable. In this case, we can use the information in the correlation table to think about what will happen if we omit the IQ variable. If IQ is omitted, it will bias the coefficient on education if two things are true: (1) there is correlation between the omitted variable (IQ) and the dependent variable (wages), and (2) there is correlation between the omitted variable (IQ) and the independent variable we are concerned with (education). If both of these are true, then our coefficient on education is likely biased. From the correlation table we can see that there is a moderate positive correlation between IQ and wages (0.325), and a positive correlation between IQ and education (0.489). Based on this, we would predict a positive bias on the estimated education coefficient if IQ is omitted. If we are unable to control for differences in IQ, our equation is likely to overstate the importance of education. To give you the idea, when I estimated the regression from part a without the IQ variable, the coefficient on education increased from 42.97 to 64.85.

e. The issue we are generally checking with residual plots is whether our assumption of homoscedastic errors has been violated. This means we are looking to see if the variance in the errors changes as we adjust the values of each x variable. When I look at this residual plot for the IQ variable, I am concerned about the variance. There appears (to me) to be an increase in the variance as IQ increases, but I think this is mostly due to some outliers that earn relatively high incomes. What we know about Jerry, specifically, is that he is below average IQ (he is around 80, and the average is around 100), and that he has a large positive residual. This means that he actually makes a lot more money each week than we predict from our equation based on his observable attributes (education, experience, age, and IQ). The residual value is around $1,000, so he makes roughly $1,000 more per week than what is predicted. Overall, it seems to be a pretty sweet life that Jerry has carved out for himself. Unfortunately, Jerry might not be bright enough to realize how truly lucky he is.

f. The first thing I notice is how much greater the coefficient on education is when we remove the IQ variable. This is an issue we discussed early in the class. IQ and education are highly correlated, so if we only put one in, it will likely overstate the impact of education. Another way of saying this – part of the impact that Model 2 estimates education has on income is coming from the fact that those with more education tend to have higher IQ (and would have theoretically made more even without the increase in education). The experience coefficient also increased, and I think this is mostly due to the fact that we removed the age variable (age and experience are correlated). Choosing the preferred model is a little trickier than it might seem. The adjusted-R^2 is higher in Model 1, but that model also likely has a collinearity problem. Removing age seems like a good move, but removing IQ causes some omitted variable bias. There is a reason this is an example that is often discussed – it is a particularly tricky one to estimate.

12. Suppose we are running a bakery and are interested in determining how many pies we will sell each week. We are going to estimate the equation $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$, where y=number of pies sold in the week, x_1 = price of pies (in dollars), and x_2 is a dummy variable equal to 1 if a holiday occurred during the week and 0 if there was no holiday.

a. The coefficient on the price variable tells us that we expect the number of pies sold to decrease by 30 every time we increase the price by $1, all else equal. The coefficient on the holiday dummy variable means that, holding price constant, we expect to sell an average of 15 more pies during a week that includes a holiday. To predict the number of pies sold when the price is $5 per pie

On non-holiday weeks: $\hat{y} = 300 - 30(5) + 15(0) = 150$

On holiday weeks: $\hat{y} = 300 - 30(5) + 15(1) = 165$

b. For sales in a holiday week (X_2=1): $\hat{y} = 340 - 10x_1$

For sales in a non-holiday week (X_2=0): $\hat{y} = 320 - 40x_1$

The interaction term is telling us how the effect of price on sales differs between holiday and non-holiday weeks. Here we find out that if we increase the price during a holiday week, we do not cause as big of a drop in sales as the same price increase would cause in a non-holiday week. In Economics terms, we would say customer demand is more inelastic during holiday weeks. To me this is reasonable; people gots to have their pie during the holidays, and so will not be as sensitive to the price. During a normal (non-holiday) week, people don't need pie unless the price is right or unless they have just had a stressful week and that phone call from two-faced Tanya was the last straw. We can use our equations above to again predict the number of pies sold in each type of week with a price of $5:

In a holiday week (X_2=1): $\hat{y} = 340 - 10(5) = 290$

In a non-holiday week (X_2=0): $\hat{y} = 320 - 40(5) = 120$

Obviously, the model with the interaction term has very different predictions for how much sales differ during holiday weeks as compared to the simple model in part a.

13. Suppose we are looking at our cross-sectional data of health and economic variables across countries from the World Bank indicators. We will estimate a model where our dependent variable is the average life expectancy in each country (in years). Our explanatory variables our: GDP per capita (in 1,000s of US$), % of the country with access to improved water, and % of young children that have been immunized for measles. Note that the percentage variables are measured on a 0 to 100 scale (instead of 0 to 1). This means if you are predicting for a country that has 50% access to clean water, you would plug in a value of 50 (instead of 0.50).
    a. The estimated coefficients tell us:
        - All else equal, if GDP per capita increases by $1,000 in a country, life expectancy will increase by .14 years.
        - All else equal, increasing access to clean water by **1 percentage point** will increase life expectancy by .36 years.
        - All else equal, increasing the measles immunization rate by **1 percentage point** will increase life expectancy by .11 years.

        What I want you to think about in terms of how much you trust the model is: (a) statistical strength (adjusted R^2, t-stats on coefficients) and (b) how concerned you are about OVB, multicollinearity, reverse causality, etc.?

    b. To predict for this country:
    *Predicted Life Expectancy = 26.27 + (0.14*12) + (0.36*95) + (0.11*60) =* ***68.75 years****.*

    c. What we are testing for here is the presence of heteroskedasticity in our errors, which would violate our OLS assumptions and lead to incorrect statistical inference (and a loss of self-esteem). The null hypothesis for this test is that there are homoskedastic errors (no problem), and the alternative is that they are heteroskedastic. When we see a p-value of (essentially) 0, it means we need to reject the null and conclude that we have a problem. This means are statistical inference tests are not to be trusted. In order to correct for this our options are to change the specification or have a software program compute robust standard errors that address the problem.

d. The standardized coefficients are helpful in that we can compare the impact of different kinds of variables on life expectancy. Does increasing GDP (improving the economy) or increasing access to clean water have a bigger impact on this health outcome? According to the standardized coefficients, a one-standard deviation increase in access to clean water has a bigger impact on life expectancy than a one-standard deviation increase in the GDP per capita in a country. These are both larger than the impact of increasing immunization against measles (by one standard deviation).

e. The coefficients on GDP and GDP squared indicate that there are positive, but diminishing, returns to increasing income in a country. This means that increasing GDP will increase life expectancy, but as the income continues to rise, the effects on life expectancy will become smaller and smaller. Note that, in this example, both the linear and quadratic coefficients are highly statistically significant. To me these results make sense. Increasing income will have a huge impact on health for low-income countries, but as incomes grow and health improves, it becomes harder to make as large of an impact. I have also seen the scatterplot, so I know what's up.

f. Guys, I screwed up. If you downloaded the practice problems right after they were posted, I asked you to predict for incomes of 12,000 and 20,000 and 30,000. This was supposed to be 10,000, 20,000 and 30,000 (which will make more sense in the discussion). I am so sorry.

   Predicted for country 1 = $29.38 + (0.36*10) - (0.0027*10^2) + (0.11*60) + (0.31*95) = 68.76$
   Predicted for country 2 = $29.38 + (0.36*20) - (0.0027*20^2) + (0.11*60) + (0.31*95) = 71.55$
   Predicted for country 3 = $29.38 + (0.36*30) - (0.0027*30^2) + (0.11*60) + (0.31*95) = 73.80$

   The idea is that the first increase in GDP by 10,000 (from GDP = 10,000 to 20,000) increased predicted life expectancy by about 2.79 years (71.55-68.76). The second increase (from 20,000 to 30,000) only increased predicted life expectancy by 2,25 years. Diminishing Returns!

g. The key to interpretations on dummy variables in recognizing the reference category (the one that is omitted). In this case, the reference category is South America. Each of the other continent coefficients are interpreted relative to South America. Life expectancy is expected to be about 9.6 years lower in an African country than a South American country, all else equal. European countries are estimated to have a higher life expectancy than South American countries (though this is not statistically significant). This model does appear to be a pretty significant improvement over the model in part a (adjusted R^2 increased from 0.677 to 0.795). Going forward, it seems we can probably just include the African continent dummy variable and be done with it, none of the others are even close to statistically significant.

14. Consider a sample of all 25 to 35-year-olds living in Idaho from the 2015 American Community Survey. We will estimate a model where we predict income (in thousands of dollars) based on various factors among those who report normally working at least 30 hours per week. We will use a logarithmic transformation of income in this model. The explanatory variables are dummy variables for female and having a bachelor's degree, and quantitative variables for age, the typical number of hours worked per week, and number of children (living in the household).

    a. Since the dependent variable has been logged, we think of the changes in income in percentage terms. This means the coefficients tell us that:
        - Females are estimated to make 31.5% less than males, all else equal
        - Those with a bachelor's degree are expected to make 43.3% more than those without a degree, all else equal
        - Each additional year of age increases income by about 0.7%, all else equal
        - Each additional child you own increases income by about 2.95%, all else equal

- Each additional hour of work per week increases income by about 1.3%, all else equal

These coefficients mostly match my expectations. There is a well-known premium for a college education, and the gender gap is a sad reality we are used to. The gap seems quite large in this one, but we haven't controlled for many things that affect income (occupation, for example). The one about kids seems a little odd, but maybe there is extra motivation to earn as much as possible when you have kids that expect to eat, have clothes, etc.

b. In the specification above, only the dependent variable has been transformed using the natural logarithm, so the question remains: should we log any of the explanatory variables? The easy part of this question is that you can rule out taking the log of the dummy variables for gender and education, as logging a dummy variable makes no sense. We could transform the age, number of children, and hours variable if we wanted to. One of the key questions we would want to consider is what will happen to our interpretations. For example, as it stands now, we are saying that each year of age increases income by a little under 1%. If we log age, then we will be interpreting as how a percentage change in age affects income. To me, it makes more intuitive sense to talk about age changes in years, rather than percentages. I feel the same way about the children and hours variables. The other question is which combination of variables produces what looks to be a linear relationship. So, we could examine the scatterplots of hours vs. ln income and ln hours vs. ln income and see which seemed to fit better with the linear assumption.

c. The main takeaway here should be that there is not a statistically significant difference in the impact of a bachelor's degree on income by gender. The results estimate that a bachelor's degree will increase income by about 40.4% for males (the coefficient on "bach"). For females, the estimated increase from a bachelor's degree is about 47.2%. This comes from taking the "bach" coefficient and adding the interaction term. As stated about however, while this seems like a large difference, there is not enough here to be statistically significant. Maybe if we had a larger sample (not just focused on Idaho), we would find significance.