Selected advanced techniques in econometrics

1. What statistical test would you use to determine if a relationship between two variables is nonlinear?

   Answer: The **Likelihood Ratio Test (LRT)**. Run the test as linear, then as quadratic, and lastly use the LRT to see if the goodness of fit (R^2) is better in the nonlinear model.

2. When should I convert the dependent variable to the **log** of that variable?

   Answer: Whenever you have a positive non-zero continuous variable. There are two advantages to doing so. First, it would typically solve the heteroskedasticity issue. Secondly, the interpretation of the coefficients is much more informative. Specifically, the coefficients would change from a level-level interpretation to a semi-elasticity interpretation. For example, suppose that you regress wage on education. If the coefficient on education is 12, then it means that for each year of education the wage per hour is likely to go up by $12. Is this a lot? Is it economically significant? Alternatively, if you regress the log wage on education, and the coefficient on education is 0.12, then it means that for each year of education, the wage is going up by 12 percent (percentage-level or semi-elasticity).

3. **Limited dependent variable**: There are many cases where the dependent variable is not a continuous variable. For example, attending a course (dummy variable), a number of times the worker was late (count data), a team a worker chose to join out of several available teams (multivariate), the worker's level of satisfaction score between 1 and 5 (ordinal). We cannot use the simple OLS regression for each of these examples. We typically make some assumptions about the distribution of the dependent variable or an underlying variable (latent variable) that determines the outcome variable, so we can estimate the model using maximum likelihood estimation. For example, if the dependent variable is binary, we can assume the probability that the outcome variable will be equal to one. It will follow a standard normal distribution (**probit** regression). That is, $\Pr(y = 1|X) = \Phi(X)$. The probability that the dependent variable (y) is equal to one conditional on some controlled variables (X) is equal to the cumulative distribution function (CDF) of a standard normal distribution (Φ). You can also assume it follows a logistic distribution (Λ), and the regression is called **logit**.

4. **Difference in differences**: Suppose you want to see if the increase in the minimum wage in Washington DC is affecting the average wages in Washington DC. So, you collected wage data in DC before the increase in min-wage and after the increase in min-wage. The

difference in the average wages cannot be directly attributed only to the increase in the minimum wage It could be that the average wage in DC would have gone up regardless of the increase in the minimum wage. So, you can take a control group, such as a neighboring state, where the minimum wage did not change. Now you compare the increase in DC to the increase in the neighboring state. If in DC the wage went up by 10% while in the neighboring state it went up by 3%, then you conclude the increase in the minimum wage in DC contributed 7% to the average wage. Technically, create a dummy that will be equal to 1 for periods after the increase in the minimum wage and 0 otherwise. Create another dummy for if your responder resides in DC and 0 otherwise. Then run a regression of log wage on the two dummies and the interaction between the two dummies. The 7% would show in the estimated coefficient of the interacted variable.

5. **Regression Discontinuity Design (RDD):** Suppose that managers at Amazon rank their workers. So, each worker has a grade between 0 and 100. Then the managers give a bonus to those who scored above 80. If you want to compare the effect of the bonus on the wage of the worker, you cannot just run a regression of the bonus on the log wage. Workers who receive the bonus are hardworking and are likely to make more than those who do not work as hard. Alternatively, you can compare the wage of a bonus recipient who scored just above 80 to those who did not receive the bonus and scored just below 80. Those two types of workers are identical in terms of their level of how hard they work. It's best to first smooth the sample by calculating the average wage by groups based on the score. For example, the average wage for workers by a point in their score. The average wage for those who scored 79, another average for those who scored 80, etc. Then, run two regressions of the average wage on the score for those who scored below 80 and another regression for those who scored above 80. The effect of the bonus is the predicted wage at score 80 in the regression above 80 to that in the regression below 80. The advantage of smoothing the sample is twofold. First, you can create an easy-to-read graph to plot the samples. Secondly, you reduce some of the variability in the sample.

6. **Synthetic control**: You would like to compare the treatment effect group to a control group. In this technique, the control group is a predicted outcome of the treatment group if untreated. For example, if you would like to test the effect of promotion on wage growth. It's not a good idea to compare those who were promoted to those who were not. The promoted ones are likely to be very motivated and more productive. So, they would have earned more regardless of the promotion. Instead, you need to compare the wage of the promoted workers to their wage if they were not promoted. You do not observe their wages if they are not promoted. So, you create a model to predict the wage of the treatment group before being promoted (using all the workers – even if not promoted). You then generate the predicted value of the promoted workers after promotion and compare these predictions to the actual wages following the promotion.

7. **A/B testing:** A hypothesis testing. You want to compare a treatment group to a control group. This testing is used extensively in tech companies such as Amazon, Google, Facebook, Uber, etc. For example, if Amazon is considering changing their website, such as changing the color from brown to purple (maybe because someone suggested that people are more likely to spend money when they see the color purple). So, you create an experiment when some users see the purple website while the others see the usual brown website. Then you compare the outcome (such as revenue or number of sales) in the treatment group to that in the control group. Whenever you run such experiments, the most important aspect is to design the experiment such that the two groups are identical except for one change (in our case, the color). There are many potential biases (i.e., reasons why the groups are not identical). For example, the novelty effect is when people (buyers) behave differently when they see something new. So, in our case, it could be that the buyers in the treatment group (purple) would initially buy more because of the novelty effect. Afterwards, they will return to their usual buying habits. One way to overcome that bias is to run the experiment long enough (more than two weeks) and only collect the data after those fourteen days. The novelty effect is likely to disappear by then.

8. **Propensity score matching (PSM):** This is a statistical technique to correct for biases of non-randomized selection into the treatment and control groups. For example, suppose that in a workplace workers have the option to take a course (on-the-job training) that will improve their productivity. As a researcher, you observe only if the worker took the course or not. You cannot see the reasons that motivated him/her to take the course. Next, you next want to evaluate the worker's productivity via a measurement that is available to you, such as the revenue generated by the worker. This will be viewed as an effect of taking the course. If you compare the productivity of workers who took the course (treatment group) to those who did not (control group), you will overestimate the effect of the course. It is likely the workers who attended the course are more motivated, committed, and productive regardless of the course. PSM is a method that assigns a score (probability) for each worker who attended the course. That is, workers who chose to attend the course will get a higher score, on average, than those who did not attend the course. Then you match each worker in the treatment group to a worker (or a group of workers) in the control group with a similar (or fairly close) score.

9. **Randomized Controlled Trial (RCT):** Suppose you want to know how the Amazon Prime subscription affects sales. You can create an experiment that **randomly** assigns Amazon users to the treatment group (those with Prime) and the others to the control group (non-Prime subscribers). If the assignments to the groups are indeed random, then you can run a simple OLS regression of the sales on a dummy variable that equals 1 if the user is in the treatment group and zero otherwise. The coefficient on the dummy is what we call the ATT (Average Treatment Effect on the Treated). In this case, it's the increase in sales,

on average, for those who have Prime compared users without Prime. ATE (Average Treatment Effect) estimates the effect of the average in the entire sample (the average effect of being a Prime subscriber for those treated and those who are untreated). In our case, we probably are more interested in ATT because we want to know the change in sales, on average, for Prime subscribers. What would you do if the trial is not randomized? In real life, it's likely that those who choose to subscribe are the ones wanting to increase their shopping. So, there is a selection bias. Specifically, the treatment group is more likely to include users who would shop more than users in the control group regardless of the Prime subscription. One way to overcome this problem is to do the PSM discussed in the previous technique. When you do the matching, you need to include a variable that can explain the selection of the treatment group. In our example, we can use the income of the user or the user's shopping habit before joining Prime.

10. How to choose a **sample size** needed for an experiment? It depends on four elements: Type I error, Type II error, Minimum Detectable Effect (MDE), and the baseline conversion rate. MDE- if we set the MDE=5%, then we will not be able to reject an effect that is smaller than 5% from zero. The MDE tells us how sensitive (or economically significant) the test should be. The baseline conversion rate is the rate of people completing the task in the control group. An example is testing whether users are making purchases with their Prime membership. The baseline conversion rate is the percentage of people purchasing with no Prime subscription.

11. How do we test for **statistical independence** between the treatment and control groups? We calculate the difference between the average values of key demographic variables (such as gender, region in the US, race, etc.). Then we test whether the differences in the means of these variables are different than zero (simple t-test).