For questions (1) through (6), you want to estimate the effect of fracking on two variables: (1) the income of households, Y, and (2) the purity of local water supplies P. Let F_{ct} be a binary variable that indicates the presence of the hydraulic fracking industry in county c at time t.

(1) Suppose you have data, (Y_{ict}, F_{ct}) for *two* periods of time, $t = 1, 2$ and for *two* counties: $c = 1, 2$.

1. Letting α be the effect of fracking on household incomes, write a *difference-in-differences* model of outcomes.

2. Without assuming any particular values of F_{ct}, write the population limit of the difference-in-differences estimator we saw in class.

3. Suppose $F_{11} = 0$, $F_{21} = 0$, $F_{12} = 1$, and $F_{22} = 1$. Do you think you would be able to estimate α in this case? Explain why not.

4. Suppose $F_{11} = 1$, $F_{21} = 0$, $F_{12} = 1$, and $F_{22} = 1$. Do you think you will be able to estimate α in this case? Explain how.

(2) Suppose you have data, (Y_{ict}, F_{ct}) for *two* periods of time, $t = 1, 2$ and for *two* counties: $c = 1, 2$. Suppose that there is only fracking in county 2 in time 2 ($F_{22} = 1$).

1. Suppose that each sample of households $(Y_{ict})_{i=1}^{N_{ct}}$ across counties and over time is independently collected, with sample sizes given by N_{ct}. Describe your estimator of α, the variance of your estimator, and how you would construct a 95% confidence interval for your estimate. Be careful to state the assumption you are using.

2. Suppose now instead that your dataset is a panel for each county, so that Y_{ict} is measured for individual i in county c at times $t = 1$ and $t = 2$ (i is now the same individual). Describe your estimator of α, the variance of your estimator, and how you would construct a 95% confidence interval for your estimate. Be careful to state the assumption you are using.

(3) Suppose now instead you have an iid dataset of observations $(Y_{ict})_{i=1}^{N}$ for many counties (C of them) over several periods of time, T. Fracking starts in different counties at different points in time across the periods $t = 1, 2, ..., T$. The total number of observations you have is N.

1. Write a linear model for household income as a function of county, c, time, t, and fracking F_{ct}, that adheres to the parallel trends assumption and for which the effect of fracking in a county on income is α.

2. Describe how you would estimate α using this dataset.

3. Describe how you would conduct a 95% significant test of the null hypothesis that $\alpha = 0$.

4. Now suppoe that we have an additional dummy variable, M_c, that indicates whether the mining industry already has a significant presence in county c. Re-write your initial model to allow the effect of fracking on income to depend on M_c (i.e. you have two effects: α_0, α_1).

5. Describe how you could estimate this new model, and how you would conduct a 95% significant test of the null hypothesis that a significant mining presence ($M_c = 1$) has *no impact* on the effect of fracking on income.

(4) Taking the same data as question (3), let t_c^0 indicate the time period in which fracking is first introduced in county c. Let $TF_{ct} = t - t_c^0$ be the number of time periods that fracking has been present in county c. Suppose now that the true effect of fracking on income changes with TF_{ct}, as:

$$\alpha(TF) = \alpha_0 + \alpha_1 TF.$$

1. Re-write your Diff-in-Diff model to allow for these time-varying effects.
2. Describe how you would estimate α_0 and α_1, and how you would conduct a 95% significant test of the null hypothesis that the impact of fracking on income is *constant over time*.

(5) Taking the same data as question (3) and (4), suppose that $C = 2$, $T > 2$, and that fracking is not introduced in either county in $t = 1, 2$. We are going to test the parallel trends assumption on these data. Suppose you estimate the model:

$$Y_{ist} = \pi_{st} + \epsilon_{ist}.$$

on just the first two periods of data (no fracking), where π_{st} is a joint time-state effect, so by definition $\pi_{st} = \mathbb{E}[Y_{ist}]$ for $s = 1, 2$ and $t = 1, 2$.

1. What does the parallel trends assumption imply about $\pi_{12} - \pi_{11}$ relative to $\pi_{22} - \pi_{21}$?
2. What does the parallel trends assumption imply about $\pi_{21} - \pi_{11}$ relative to $\pi_{22} - \pi_{12}$?
3. Describe how you would conduct a *joint hypothesis test* of each of the above to implications. Supposing you reject the null hypothesis, what does it imply about the parallel trends assumption and the work we have done so far?
4. Suppose that $C = 3$. Propose at least two further restrictions that you could add to the joint test, and how to incorporate them into your test above.

(6) Now, supposing that you have many counties, C, over several periods, T, with F_{ct} given for each combination of c and t.

1. Describe how you would estimate the impact, κ, of fracking on water purity, P_{ct}, if you also have these data for each pair (c, t).
2. Describe how you would approximate (i.e. estimate) the variance of your estimator.
3. Supposing that an independent analyst has claimed that the total economic value of water purity is $\$q$ per person, per unit of measurement used to construct P. Assuming this number is correct, describe how you would construct a confidence interval for the *total economic benefit* of fracking: $\alpha + \kappa q$[1].

(7) Suppose that you have data from a randomized control trial, where Y is the treatment outcome of interest. Suppose that you have *pre-treatment* data for both treatment and control groups, $(Y_{i0T})_{i=1}^{N_T}$, $T = 0, 1$, and post-treatment data $(Y_{i1T})_{i=1}^{N_T}$. Only the treatment group ($T = 1$) in the post-treatment period have received the treatment, while the control group receives a *placebo*. Let $t = 0, 1$ indicate the pre and post-treatment periods.

[1] You may note that κ is most likely to be negative

Suppose that the total effect of the treatment, α, includes a *placebo effect*, α_0, and a *true effect*, α_1. This gives that $\alpha = \alpha_0 + \alpha_1$. The experimental model of outcomes is:

$$Y_{itT} = \mu + \alpha_0 D_t + \alpha_1 D_t T_i$$

where D_t is a dummy variable equal to 1 for observations in the post-treatment period ($t = 1$).

1. Show you could use the pre and post-treatment outcomes for the control group to estimate the placebo effect, α_0.
2. Show how you could use difference-in-differences to estimate the true effect, α_1.
3. Show how you could use the pre-treatment data on outcomes Y to test that the treatment and control groups are comparable (you may use any significance you like for this test).

(8) State the necessary conditions for two-stage-least squares to be consistent and asymptotically normal. Which condition would I need to additionally assume in order to derive the following variance formula?

$$\mathbb{V}[\hat{\beta}_{2SLS}] = \frac{\sigma^2}{N}(\mathbf{Q}_{XZ}\mathbf{Q}_{ZZ}^{-1}\mathbf{Q}_{XZ}')^{-1}$$

How would I estimate the variance if I was unwilling to make this additional assumption?

(9) Suppose that the labor supply of individual i in county c can be described as:

$$\log(h_{ic}) = \psi_0 + \psi_1 \log(w_{ic}) + \alpha_i$$

where h_{ic} is hours of work, and w_{ic} is their hourly wage. Second, assume that wages can be described as:

$$\log(w_{ic}) = X_{ic}\beta + \gamma U_c + \mu_i$$

where X_{ic} is a vector of demographics (including a constant), U_c is the county-level unemployment rate (a proxy for labor demand), α_i represents person-level differences in preferences for work, and μ_i represents person-level differences in ability. We can normalize $\mathbb{E}[\alpha_i] = \mathbb{E}[\mu_i] = 0$, but you expect that $\mathbb{C}(\mu_i, \alpha_i) > 0$.

1. Supposing we had iid data on hours h_{ic}, and wages w_{ic} for individuals across counties, do you expect to be able to consistently estimate the parameters ψ_0, ψ_1 by OLS? Why or why not? What direction will the bias go in?
2. Now suppose that you additionally have data X_{ic} and U_c. Suppose that individual traits $(\mu_i, \alpha_i, X_{ic})$ are independant of the unemployment rate U_c. How can you now estimate ψ? Do you need to include the data X_{ic} in this process at all? Why or why not?
3. Describe how you would test the hypothesis that labor supply is perfectly inelastic.

(10) Taking the same setup as above, suppose that our data is collected over different time periods. The model can be described as:

$$\log(h_{ict}) = \psi_0 + \psi_1 \log(w_{ict}) + \alpha_i + \epsilon_{ict}$$
$$\log(w_{ict}) = \phi_c + \gamma U_{ct} + \mu_i + \zeta_{ict}$$

where t indexes time, and we are dropping any dependance of the wage on observables, X_{ic}, but allowing for county-specific effects on wages, ϕ_c. You can assume that the new error terms ϵ_{ict} and ζ_{ict} are independent of everything else.

1. Suppose you are worried that individuals' preferences for work are higher in counties with higher wages (ϕ_c). This would be true, for example, if high skill industries concentrate in particular locations. Express this concern in terms of the relationship between ϕ_c and α_i.

2. What condition on unemployment, U_{ct}, would guarantee that the IV estimator from the previous question is still consistent, even if the above relationship holds?

3. Suppose that this condition is violated, what weaker assumption could we make, and how would we implement the IV estimator? Hint: suppose we can write unemployment as

$$U_{ct} = \varphi_c + \xi_{ct}$$

and use a condition on ξ_{ct}.

(11) Suppose you have L instruments, Z_i, for a single variable, x_i. In the first stage you estimate:

$$x_i = \pi_0 + Z_i\pi_1 + \eta_i$$

where π_1 is a $L \times 1$ vectore of coefficients. You are worried that your instruments might not actually have a significant impact on x_i. Propose a test of the joint hypothesis that each coefficient of π_1, $\pi_{1,l} = 0$ for $l = 1, 2, ..., L$. If you reject the null hypothesis, what practical information does this offer about your chosen instruments?

(12) Suppose you are estimating the model

$$Y_i = \beta_0 + \beta_1 H_i + \beta_2 C_i + \epsilon_i$$

where Y_i is income, H_i is a dummy indicating if the individual has graduated from high school, and C_i a dummy that indicates graduation from college. You want to interpret β_1 and β_2 as causal effects of these education variables. To help, you have two instruments, $Z_{1,i}$ is an instrument based on college openings, and $Z_{2,i}$ an instrument based on eligibility for a college tuition subsidy. The first stage can be written as:

$$H_i = \pi_0 + \pi_1 Z_{1,i} + \pi_2 Z_{2,i}$$
$$C_i = \gamma_0 + \gamma_1 Z_{1,i} + \gamma_2 Z_{2,i}$$

1. Since both your instruments are related to college, your are worried that $\pi_1 = \pi_2 = 0$. Which assumption would be violated if this were true?

2. How could you test to see if this will be an issue?

3. Suppose that you reject the null hypothesis in the previous test. Provide some economic intuition for why these college-focused instruments might also affect high school.

(13) Consider the following model:

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

where $\mathbb{E}[\epsilon_i | x_i] = 0$, and your variable of interest, x_i, is measured with error:

$$x_i^m = x_i + \eta_i$$

and η_i is independent of all other variables.

1. Explain why OLS using Y_i and x_i^m would produce an inconsistent estimator for β_1.
2. Suppose you had a second measure of x_i, z_i:

$$z_i = x_i + \zeta_i$$

where ζ_i is independent of all other variables. Show how you can estimate β_1.