The questions below are slightly more difficult. Don't panic if you struggle with them, the are a bit above exam-level difficulty, but doing them will help you prepare for the exam and learn this content.

1) In this question we are going to solve a generalized version of the selection model we saw in class. Suppose that utility from the college decision d is given, for individual i, by:

$$U_i(d) = (1 - d)\gamma_0 + d(\gamma_1 + u_i)$$

where u_i is distributed as a *standard normal random variable*. Thus, the individual chooses college $(D_i = 1)$ if

$$u_i > \gamma_0 - \gamma_1.$$

Suppose that potential wages $W_{D,i}$ for $D = 0, 1$ can be written as

$$W_{D,i} = \mu_D + \eta_{D,i}$$

where $\mathbb{E}[\eta_{D,i}] = 0$. In addition, assume that the pair of error terms $\eta_{1,i}, \eta_{0,i}$ are *independent* of each other, but each is *not* independent of u_i. Suppose that $\mathbb{C}(\eta_{1,i}, u_i) = \sigma_{u1}$, $\mathbb{C}(\eta_{0,i}, u_i) = \sigma_{u0}$, and $\mathbb{V}[\eta_{D,i}] = \sigma_D^2$ for $D = 0, 1$.

1. What is the average treatment effect of college attendance on earnings?

2. Suppose you observe wages, W_i, and college choice, D_i, for an iid sample of individuals. Calculate $\mathbb{E}[W_i | D_i = d]$ for $d = 0, 1$. You may use the following fact: if Z is a standard normal, and X is normal with $\mathbb{V}[X] = \sigma_X^2$, $\mathbb{C}(X, Z) = \sigma_{XZ}$, and $\mathbb{E}[X] = 0$:

$$\mathbb{E}[X | Z > c] = \frac{\sigma_{XZ}}{\sigma_X} \frac{\phi(c)}{1 - \Phi(c)}, \qquad \mathbb{E}[X | Z < c] = -\frac{\sigma_{XZ}}{\sigma_X} \frac{\phi(c)}{\Phi(c)}$$

where Φ is the cdf of the standard normal and ϕ is the pdf.

3. Suppose I take a sample mean of wages for college workers, $\overline{W}_1$, what is $\mathbb{E}[\overline{W}_1]$? What will the sample mean converge to?

4. Suppose I take a sample mean of wages for non-college workers, $\overline{W}_0$, what is $\mathbb{E}[\overline{W}_0]$? What will the sample mean converge to?

5. What will $\overline{W}_1 - \overline{W}_0$ converge to? Will this in general be equal to the average treatment effect?

6. Under what restrictions on the parameters $\sigma_{1u}, \sigma_{0u}, \mu_1, \mu_0$ will the $\overline{W}_1 - \overline{W}_0$ converge to the ATE?

7. Suppose $\sigma_{1u} > 0$ and $\sigma_{0u} > 0$. If we used $\overline{W}_1 - \overline{W}_0$ as an estimate for the ATE, would we over or under-estimate the ATE?

2) Suppose that you are evaluating the effect of a treatment, T_i, on the outcomes of children. However, the treatment allocation is clustered at the family level, so outcomes for child i in family f can be written as:

$$Y_{fi} = \alpha_0 + \alpha_1 T_f + \epsilon_{fi}$$

You can assume that the experiment has been properly run and there are no issues with selection. Suppose that you assign N_F families the treatment, N_F families are assigned to the control group, with two children in every family, so $N = 2N_F$ for each group. Assume that families are sampled iid, $\mathbb{V}[\epsilon_{fi}] = \sigma_\epsilon^2$ and $\mathbb{C}(\epsilon_{f1}, \epsilon_{f2}) = \sigma_{12}$ (i.e. error terms for children within the same family are not independent).

1. Propose an estimator for the average treatment effect. Is this estimator unbiased?

2. What is the problem with the typical variance formula?

3. Define $Y_f = Y_{f1} + Y_{f2}$. Compute $\mathbb{V}[Y_f]$.

4. Use the fact that Y_f is independent across families to compute $\mathbb{V}[\sum_{f=1}^{N_F} Y_f]$ for all the families in one group.

5. Noting that the sample mean for each group can be written as $\frac{1}{N} \sum_{f=1}^{N_F} (Y_{1f} + Y_{2f})$, calculate the variance of $\overline{Y}_1$ and $\overline{Y}_0$, the sample means from each group.

6. Use your answer above to derive the asymptotic distribution of your estimator for the average treatment effect, and describe how you would estimate the variance of this distribution.

7. How does the variance of your estimator compare to the case in which the treatment is assigned randomly to individual children, and only one child per family is chosen for the sample.