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? ATE is $\mathbb{E}[W_{1,i}] - \mathbb{E}[W_{0,i}] = \mu_1 - \mu_0$
- 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?

It will converge to

$$\mathbb{E}[W_{1,i}|u_i > \gamma_0 - \gamma_1] = \mu_1 + \mathbb{E}[\eta_{1,i}|u_i > \gamma_0 - \gamma_1] = \mu_1 + \frac{\sigma_{1u}}{\sigma_1} \frac{\phi(\mu_0 - \mu_1)}{1 - \Phi(\mu_0 - \mu_1)}$$

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? It will converge to

$$\mathbb{E}[W_{0,i}|u_i < \gamma_0 - \gamma_1] = \mu_0 + \mathbb{E}[\eta_{0,i}|u_i < \gamma_0 - \gamma_1] = \mu_0 - \frac{\sigma_{0u}}{\sigma_0} \frac{\phi(\mu_0 - \mu_1)}{\Phi(\mu_0 - \mu_1)}$$

5. What will $\overline{W}_1 - \overline{W}_0$ converge to? Will this in general be equal to the average treatment effect? Putting together the two answers above, it will converge to:

$$\mu_{1} - \mu_{0} + \underbrace{\frac{\sigma_{1u}}{\sigma_{1}} \frac{\phi(\mu_{0} - \mu_{1})}{1 - \Phi(\mu_{0} - \mu_{1})}}_{\text{Bias}} + \frac{\sigma_{0u}}{\sigma_{0}} \frac{\phi(\mu_{0} - \mu_{1})}{\Phi(\mu_{0} - \mu_{1})}$$

In general, since $\phi > 0$ and $\Phi > 0$, the terms involving the truncated expectation will not be equal to zero, and so the difference in sample means will not converge to the ATE.

- 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? Based on the above analysis, if $\sigma_{1u} = \sigma_{0u} = 0$, then the bias term above is equal to zero, and we can recover the ATE. This is the case in which the decision to attend college is *independent* of potential wages, $W_{1,i}, W_{0,i}$ (hence there is no selection). Another interesting case is when the ATE is zero, i.e. $\mu_1 - \mu_0 = 0$, and $\frac{\sigma_{1u}}{\sigma_1^2} = \frac{\sigma_{0u}}{\sigma_0^2}$. In this case, we would have selection, but the parameters are such that the selection bias on each population mean exactly cancels out.
- 7. Suppose σ_{1u} > 0 and σ_{0u} > 0. If we used W
 ₁ W
 ₀ as an estimate for the ATE, would we over or under-estimate the ATE?
 If both covariances are positive, then the bias term must be positive since all other terms in the equation are positive by definition. We would therefore over-estimate the ATE.

2) Suppose that you are evaluating the effect of a treatment, T, 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? Use the difference in sample means: $\hat{\alpha} = \overline{Y}_1 - \overline{Y}_0$. Note:

$$\mathbb{E}[\overline{Y}_1] = \mathbb{E}[\frac{1}{N}\sum_{f=1}^{N_F}\sum_{i=1,2}Y_{if}] = \frac{1}{N}N\mathbb{E}[Y_{if}] = \alpha_0 + \alpha_1$$

similar logic for the sample of control families gives that $\mathbb{E}[\hat{\alpha}] = \alpha$.

- 2. What is the problem with the typical variance formula? The data are not iid because of the correlation across children within families.
- 3. Define $Y_f = Y_{f1} + Y_{f2}$. Compute $\mathbb{V}[Y_f]$. $\mathbb{V}[Y_{f1} + Y_{f2}] = 2\sigma_{\epsilon}^2 + 2\sigma_{12}$
- 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. Since the Y_f are independent, $\mathbb{V}[\sum_{f=1}^{N_F} Y_f] = N_f \mathbb{V}[Y_f] = 2N_f(\sigma_{\epsilon}^2 + \sigma_{12})$
- 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.

$$\mathbb{V}[\overline{Y}_0] = \mathbb{V}[\overline{Y}_1] = \mathbb{V}\left[\frac{1}{N}\sum_{f=1}^{N_F} (Y_{1f} + Y_{2f})\right]$$
$$= \frac{1}{N^2} \times 2N_f(\sigma_\epsilon^2 + \sigma_{12})$$
$$= \frac{\sigma_\epsilon^2 + \sigma_{12}}{N}$$

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. Since we have assumed here that the variances and covariances of ϵ_{fi} are constant in both groups, we get that:

$$\mathbb{V}[\hat{\alpha}] = \mathbb{V}[\overline{Y}_1 - \overline{Y}_0] = 2\frac{\sigma_{\epsilon}^2 + \sigma_{12}}{N} = \frac{\sigma_{\epsilon}^2 + \sigma_{12}}{N_F}$$

We can estimate the variance by calculating the sample variance of $Y_{f1} + Y_{f2}$ from each group and substituting this into the expression above. Our work has shown that $\hat{\alpha}$ should be asymptotically normal with mean α and variance given above.

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.

In this case the variance of the estimator would simply be $\frac{\sigma_{\epsilon}^2}{N}$. If $\sigma_{12} > 0$, then the estimator from this sample would be more efficient. However if $\sigma_{12} < 0$ and sufficiently close in magnitude to σ_{ϵ}^2 , then it is possible for the estimator using the sample above to be more efficient.