ECMA 31100: Problem Set 1

Due January 29 by 11:59PM

Question 1 You observe wages (y_i), whether an individual took job training ($d_i \in \{0,1\}$), and a scalar covariate x_i measuring prior labor market outcomes in an iid sample of n individuals. Define the potential outcomes y_{i0}, y_{i1} as

$$\begin{aligned} y_{i0} &= \text{wage of individual } i \text{ without job training}, \\ y_{i1} &= \text{wage of individual } i \text{ with job training}. \end{aligned}$$

You are interested in studying the average treatment effect, $\mathrm{E}\left(y_{i1} - y_{i0}\right)$. Assume that d_i is randomly assigned conditional on x_i. That is:

$$(y_{i0}, y_{i1}) \perp\!\!\!\perp d_i | x_i.$$

a) How plausible is the assumption that $y_{i1} - y_{i0}$ is constant across individuals?

b) How plausible are the assumptions $(y_{i0}, y_{i1}) \perp\!\!\!\perp d_i$ and $(y_{i0}, y_{i1}) \perp\!\!\!\perp d_i | x_i$?

Suppose in addition that $\mathrm{E}\left(y_{i1} - y_{i0} | x_i\right) = c$, for some constant c.

ci) Is the assumption $\mathrm{E}\left(y_{i1} - y_{i0} | x_i\right) = c$ stronger or weaker than assuming $y_{i1} - y_{i0} = c$?

cii) Show that $\mathrm{E}\left(y_i | d_i = 1, x_i\right) - \mathrm{E}\left(y_i | d_i = 0, x_i\right) = c$.

Suppose in addition that $\mathrm{E}\left(y_{i0} | x_i\right) = \beta_0 + \beta_1 x_i$ for some unknown constants β_0, β_1.

d) Show that we may write

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 d_i + e_i, \tag{1}$$

where $\mathrm{E}\left(e_i | d_i, x_i\right) = 0$. Show that β_2 equals the average treatment effect.

Now drop the assumption that $\mathrm{E}\left(y_{i1} - y_{i0} | x_i\right)$ is constant, but instead assume that x_i takes two values $x_i \in \{0,1\}$.

e) Argue that the representation in part d) is not necessarily valid, but the following representation is valid:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 d_i + \beta_3 d_i x_i + v_i; \quad \mathrm{E}\left(v_i | d_i, x_i\right) = 0. \tag{2}$$

What is β_2? Show that β_3 is a difference between (conditional) average treatment effects.

f) Suppose you estimate the coefficients in the regression

$$y_i = b_0 + b_1 x_i + b_2 d_i + e_i,$$

using OLS. Show that $\hat{b}_2$ converges to a weighted average of conditional average treatment effects, and generally not the ATE.

ANS: a) The training program may have little effect on people who already have the skills it teaches, and a much greater effect on those who don't. The participants may also respond differently to the program depending on their learning style, so the constant effects assumption is hard to justify.

b) A training program may have differential effects on workers with different prior labor market outcomes. If prior labor market outcomes are predictive of future earnings, and those with better prior labor market outcomes derive less benefit from the program, d_i may be positively correlated with $y_1 - y_0$. On the other hand, if we fix prior outcomes, participants may benefit similarly. So, we might believe that entry into the program is as good as randomly assigned conditional on prior labor market outcomes. Of course, you may not believe this either!

ci) Weaker. If $y_{i1} - y_{i0}$ is constant, its (conditional) expectation is constant, but a constant conditional expectation doesn't imply a random variable is constant.

cii) We have

$$\begin{aligned} \mathrm{E}\left(y_i|d_i = 1, x_i\right) - \mathrm{E}\left(y_i|d_i = 0, x_i\right) &= \mathrm{E}\left(y_{i1}|d_i = 1, x_i\right) - \mathrm{E}\left(y_{i0}|d_i = 0, x_i\right) \\ &= \mathrm{E}\left(y_{i1}|x_i\right) - \mathrm{E}\left(y_{i0}|x_i\right) \\ &= \mathrm{E}\left(y_{i1} - y_{i0}|x_i\right) = c, \end{aligned}$$

where the first equality follows because $y_i = y_{i0} + d_i\left(y_{i1} - y_{i0}\right)$, the second because of unconfoundedness, and the final equality is by assumption.

Suppose in addition that $\mathrm{E}\left(y_{i0}|x_i\right) = \beta_0 + \beta_1 x_i$ for some unknown constants β_0, β_1.

d) Using part c), we have $\mathrm{E}\left(y_{i1}|x_i\right) = \beta_0 + \beta_1 x_i + c$. Therefore,

$$\mathrm{E}\left(y_i|d_i, x_i\right) = \beta_0 + \beta_1 x_i + c d_i,$$

from which the representation in (1) follows with $\beta_2 = c$. Since $c = \mathrm{E}\left(y_{i1} - y_{i0}|x_i\right)$, the law of iterated expectation implies $c = \mathrm{E}\left(y_{i1} - y_{i0}\right)$, which is the ATE.

e) The representation in d) is no longer valid because now

$$\mathrm{E}\left(y_i|d_i, x_i\right) = \beta_0 + \beta_1 x_i + \mathrm{E}\left(y_{i1} - y_{i0}|x_i\right) d_i.$$

Since we have dropped the assumption of a constant conditional average treatment effect, $\mathrm{E}\left(y_{i1} - y_{i0}|x_i\right)$ may depend on x_i. However, since x_i takes only two values, $\mathrm{E}\left(y_{i1} - y_{i0}|x_i\right)$ may take at most two values, so WLOG, for some constants β_2, β_3,

$$\mathrm{E}\left(y_{i1} - y_{i0}|x_i\right) = \beta_2 + \beta_3 x_i,$$

which yields (2). Next, we have the following equalities:

$$\begin{aligned}
\beta_0 &= \mathrm{E}\left(y_i | d_i = 0, x_i = 0\right), \\
\beta_0 + \beta_1 &= \mathrm{E}\left(y_i | d_i = 0, x_i = 1\right), \\
\beta_0 + \beta_2 &= \mathrm{E}\left(y_i | d_i = 1, x_i = 0\right), \\
\beta_0 + \beta_1 + \beta_2 + \beta_3 &= \mathrm{E}\left(y_i | d_i = 1, x_i = 1\right).
\end{aligned}$$

Noting that $y_i = y_{i0} + d_i\left(y_{i1} - y_{i0}\right)$ and solving gives

$$\begin{aligned}
\beta_2 &= \mathrm{E}\left(y_{i1} | d_i = 1, x_i = 0\right) - \mathrm{E}\left(y_{i0} | d_i = 0, x_i = 0\right) \\
&= \mathrm{E}\left(y_{i1} - y_{i0} | x_i = 0\right),
\end{aligned}$$

where the second equality follows because of unconfoundedness. Similarly,

$$\begin{aligned}
\beta_3 &= \mathrm{E}\left(y_i | d_i = 1, x_i = 1\right) - \mathrm{E}\left(y_i | d_i = 0, x_i = 1\right) - \beta_2 \\
&= \mathrm{E}\left(y_{i1} - y_{i0} | x_i = 1\right) - \beta_2 \\
&= \mathrm{E}\left(y_{i1} - y_{i0} | x_i = 1\right) - \mathrm{E}\left(y_{i1} - y_{i0} | x_i = 0\right),
\end{aligned}$$

which is a difference between average treatment effects conditional on $x_i = 1$ and $x_i = 0$. $\beta_3 = 0$ in d) because there we assumed a constant conditional average treatment effect.

f) First note that since $\beta_2 = \mathrm{E}\left(y_{i1} - y_{i0} | x_i = 0\right)$, we can use partitioned regression to derive b_2:

$$b_2 = \frac{\mathrm{E}\left(\tilde{d}_i y_i\right)}{\mathrm{E}\left(\tilde{d}_i^2\right)} = \beta_2 + \beta_3 \frac{\mathrm{E}\left(\tilde{d}_i \cdot d_i x_i\right)}{\mathrm{E}\left(\tilde{d}_i^2\right)},$$

where

$$\tilde{d}_i = d_i - \mathrm{P}\left(d_i = 1 | x_i\right)$$

is the residual of a (population) regression of x on d. It follows that

$$\begin{aligned}
\frac{\mathrm{E}\left(\tilde{d}_i \cdot d_i x_i\right)}{\mathrm{E}\left(\tilde{d}_i^2\right)} &= \frac{\mathrm{E}\left(x_i d_i^2 - d_i x_i \mathrm{P}\left(d_i = 1 | x_i\right)\right)}{\mathrm{E}\left(Var\left(d_i | x_i\right)\right)} \\
&= \frac{\mathrm{E}\left(d_i x_i \mathrm{P}\left(d_i = 0 | x_i\right)\right)}{\mathrm{E}\left(Var\left(d_i | x_i\right)\right)} \\
&= \frac{\mathrm{P}\left(d_i = 1 | x_i = 1\right) \mathrm{P}\left(d_i = 0 | x_i = 1\right) \mathrm{P}\left(x_i = 1\right)}{\mathrm{E}\left(Var\left(d_i | x_i\right)\right)} \\
&= \frac{Var\left(d_i | x_i = 1\right) \mathrm{P}\left(x_i = 1\right)}{\mathrm{E}\left(Var\left(d_i | x_i\right)\right)}.
\end{aligned}$$

It follows that

$$\begin{aligned}
b_2 &= \beta_2 + \beta_3 \frac{Var\left(d_i|x_i=1\right)\mathrm{P}\left(x_i=1\right)}{\mathrm{E}\left(Var\left(d_i|x_i\right)\right)} \\
&= \mathrm{E}\left(y_{i1}-y_{i0}|x_i=0\right) \\
&+ \frac{Var\left(d_i|x_i=1\right)\mathrm{P}\left(x_i=1\right)}{\mathrm{E}\left(Var\left(d_i|x_i\right)\right)}\left[\mathrm{E}\left(y_{i1}-y_{i0}|x_i=1\right)-\mathrm{E}\left(y_{i1}-y_{i0}|x_i=0\right)\right] \\
&= \frac{Var\left(d_i|x_i=1\right)\mathrm{P}\left(x_i=1\right)}{\mathrm{E}\left(Var\left(d_i|x_i\right)\right)}\mathrm{E}\left(y_{i1}-y_{i0}|x_i=1\right)+\frac{Var\left(d_i|x_i=0\right)\mathrm{P}\left(x_i=0\right)}{\mathrm{E}\left(Var\left(d_i|x_i\right)\right)}\mathrm{E}\left(y_{i1}-y_{i0}|x_i=0\right).
\end{aligned}$$

We see that the weights are non-negative and sum to one. Note that the latter quantity would be equal to the ATE if the conditional average treatment effects were equal, but it would also equal the ATE if

$$Var\left(d_i|x_i=1\right)=Var\left(d_i|x_i=0\right)=\mathrm{E}\left(Var\left(d_i|x_i\right)\right).$$

This last condition would hold, for example, if the probability of entering training conditional on prior labor market outcomes does not vary with prior outcomes, but that is equivalent (in this binary case) to independence between d_i and x_i. Note that when neither of these conditions hold, in general, the weighted average will not equal the ATE.

Question 2 Consider the potential outcomes framework

$$Y = Dy_1 + (1-D)\,y_0$$

and suppose that $(y_0, y_1) \perp D|X$, where X is a vector of covariates. Define $ATE\left(x\right) := \mathrm{E}\left(y_1 - y_0|X=x\right)$. You are given an iid sample of $\{Y_i, D_i, X_i\}_{i=1}^N$. Assume that all relevant moments exist.

a) In class we provided an identification argument for the ATE. Do the same for $ATE\left(x\right)$. Use this to provide an identification argument for the ATT and ATU.

ANS: We have

$$\begin{aligned}
ATE\left(x\right) &= \mathrm{E}\left(y_1-y_0|X=x\right) \\
&= \mathrm{E}\left(y_1|D=1,X=x\right)-\mathrm{E}\left(y_0|D=0,X=x\right) \\
&= \mathrm{E}\left(Y|D=1,X=x\right)-\mathrm{E}\left(Y|D=0,X=x\right).
\end{aligned}$$

Now

$$
\begin{aligned}
ATT &= \mathrm{E}\left(y_1 - y_0 | D = 1\right) \\
&= \mathrm{E}\left(\mathrm{E}\left(y_1 - y_0 | D = 1, X\right) | D = 1\right) \\
&= \mathrm{E}\left(ATE\left(X\right) | D = 1\right) \\
ATE &= \mathrm{E}\left(\mathrm{E}\left(ATE\left(X\right)\right)\right) \\
ATU &= \mathrm{E}\left(y_1 - y_0 | D = 0\right) \\
&= \mathrm{E}\left(\mathrm{E}\left(y_1 - y_0 | D = 0, X\right) | D = 0\right) \\
&= \mathrm{E}\left(ATE\left(X\right) | D = 0\right).
\end{aligned}
$$

From now on assume $\mathrm{E}\left(y_d | X = x\right) = \alpha_d + x'\beta_d$ for $d = 0, 1$.

b) How do your answers to part a) change?

ANS: Now we have

$$
\begin{aligned}
ATE\left(x\right) &= \mathrm{E}\left(y_1 - y_0 | X = x\right) \\
&= \alpha_1 - \alpha_0 + x'\left(\beta_1 - \beta_0\right),
\end{aligned}
$$

so

$$
\begin{aligned}
ATE &= \alpha_1 - \alpha_0 + \mathrm{E}\left(X'\right)\left(\beta_1 - \beta_0\right) \\
ATT &= \alpha_1 - \alpha_0 + \mathrm{E}\left(X' | D = 1\right)\left(\beta_1 - \beta_0\right) \\
ATU &= \alpha_1 - \alpha_0 + \mathrm{E}\left(X' | D = 0\right)\left(\beta_1 - \beta_0\right)
\end{aligned}
$$

c) Write $\mathrm{E}\left(Y | D, X\right)$ in terms of $\alpha_0, \alpha_1, \beta_0, \beta_1, D, X$. Propose consistent estimators of the ATE, ATT and ATU and prove their consistency.

ANS: We have

$$
\begin{aligned}
\mathrm{E}\left(Y | D, X\right) &= \mathrm{E}\left(y_0 + D\left(y_1 - y_0\right) | D, X\right) \\
&= \mathrm{E}\left(y_0 | X\right) + D\mathrm{E}\left(y_1 - y_0 | X\right) \\
&= \alpha_0 + \left(\alpha_1 - \alpha_0\right) D + X'\beta_0 + D \cdot X'\left(\beta_1 - \beta_0\right).
\end{aligned}
$$

It follows that in a regression of Y on a constant, $D, X, D \cdot X$ the error term has mean zero conditional on D, X which is sufficient (along with random sampling) for each of the parameter estimates to be unbiased and consistent. By combining elements in a vector and using the CMT we get that $[\widehat{\alpha_1 - \alpha_0}] + \bar{X}_n \cdot [\widehat{\beta_1 - \beta_0}]$is a consistent estimate of the ATE. Replace $\bar{X}_n$ with

$$
\bar{X}_{1,n} = \frac{\sum X_i \mathbf{1}\left(D_i = 1\right)}{\sum \mathbf{1}\left(D_i = 1\right)} \text{ or } \bar{X}_{0,n} = \frac{\sum X_i \mathbf{1}\left(D_i = 0\right)}{\sum \mathbf{1}\left(D_i = 0\right)}
$$

for consistent estimates of the ATT or ATU, respectively.

d) Show that

$$\mathrm{E}(Y|D,X) = \gamma_0 + \gamma_1 D + X'\gamma_2 + D \cdot (X - \mathrm{E}(X))' \gamma_3$$

for some values of $\gamma_0, \gamma_1, \gamma_2, \gamma_3$ you derive, and show that γ_1 represents the ATE. How would you modify the above construction to obtain γ_1 as the ATT or ATU?

ANS: We have

$$\begin{aligned}
\mathrm{E}(Y|D,X) &= \alpha_0 + (\alpha_1 - \alpha_0) D + X'\beta_0 + D \cdot X' (\beta_1 - \beta_0) \\
&= \alpha_0 + (\alpha_1 - \alpha_0) D + X'\beta_0 + D \cdot [X - \mathrm{E}(X)]' (\beta_1 - \beta_0) \\
&+ D \cdot \mathrm{E}(X)' (\beta_1 - \beta_0) \\
&= \gamma_0 + \gamma_1 D + X'\gamma_2 + D \cdot (X - \mathrm{E}(X))' \gamma_3
\end{aligned}$$

where $\gamma_0 = \alpha_0, \gamma_1 = \alpha_1 - \alpha_0 + \mathrm{E}(X)'(\beta_1 - \beta_0), \gamma_2 = \beta_0, \gamma_3 = \beta_1 - \beta_0$. Observe that γ_1 is the ATE derive in c). To obtain the ATT, instead write

$$\mathrm{E}(Y|D,X) = \gamma_0 + \gamma_1 D + X'\gamma_2 + D \cdot (X - \mathrm{E}(X|D=1))' \gamma_3$$

so that $\gamma_1 = \alpha_1 - \alpha_0 + \mathrm{E}(X'|D=1)(\beta_1 - \beta_0)$; analogously for the ATU.

e) How might you estimate γ_1 directly? You will need to replace $\mathrm{E}(X)$. Show that your estimator of γ_1 is consistent.

ANS: Regress Y_i on a constant, D_i, X_i, $D_i \cdot \left(X_i - \bar{X}_n\right)$ and use the coefficient estimate on D_i. Some algebra shows

$$\hat{\gamma}_1 = [\widehat{\alpha_1 - \alpha_0}] + \bar{X}_n \cdot [\widehat{\beta_1 - \beta_0}],$$

which is consistent because the OLS estimates are consistent, by part c).

f) Derive $Var(Y|D,X)$ in terms of $Var(y_0|X), Var(y_1|X)$ and D. Is it reasonable to assume homoskedasticity (i.e. that $Var(Y|D,X)$ is constant)?

ANS: We have

$$\begin{aligned}
Var(Y|D,X) &= Var(Y|D=1,X)\,D + Var(Y|D=0,X)(1-D) \\
&= Var(y_1|X)\,D + Var(y_0|X)(1-D).
\end{aligned}$$

Not reasonable: Heteroskedasticity can arise if the potential outcomes have different variances, even if those variances are assumed to be constant functions of X. In the job training example, we may expect the variation in y_1 to be greater than in y_0 if the program creates more employment opportunities.

g) Suppose $\mathrm{E}(X)$ is known, so that it is in fact possible to run the regression

$$Y_i = W_i'\gamma + \epsilon_i; \quad \mathrm{E}(\epsilon_i|W_i) = 0,$$

where $W_i = \left(1, D_i, X_i', D_i \cdot [X_i - \mathrm{E}(X)]'\right)'$. Show that

$$\sqrt{n}\left(\hat{\gamma}_{OLS} - \gamma\right) \xrightarrow{d} \mathcal{N}\left(0, V\right),$$

for some matrix V you specify. Provide a consistent estimate of V under homoskedasticity and another under heteroskedasticity.

ANS: We have

$$\begin{aligned}
\sqrt{n}\left(\hat{\gamma}_{OLS} - \gamma\right) &= \left(\frac{1}{n}\sum W_i W_i'\right)^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} W_i \epsilon_i \\
&\xrightarrow{d} \mathcal{N}\left(0, \mathrm{E}\left(W_i W_i'\right)^{-1} \mathrm{E}\left(\epsilon_i^2 W_i W_i'\right) \mathrm{E}\left(W_i W_i'\right)^{-1}\right).
\end{aligned}$$

Under homoskedasticity this simplifies to $\sigma^2 \mathrm{E}\left(W_i W_i'\right)$, where $\sigma^2 = \mathrm{E}\left(\epsilon_i^2\right)$. A consistent estimate under heteroskedasticity is

$$\left[\frac{W'W}{n}\right]^{-1} \left(\frac{1}{n}\sum_{i=1}^{n} \hat{\epsilon}_i^2 W_i W_i'\right) \left[\frac{W'W}{n}\right]^{-1},$$

where $\hat{\epsilon}_i$ denotes the i-th OLS residual. A simpler consistent estimate can be used under homoskedasticity:

$$\hat{\sigma}^2 \left[\frac{W'W}{n}\right]^{-1}; \quad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} \hat{\epsilon}_i^2.$$

h) Show that when $\mathrm{E}(X)$ is replaced by $\bar{X}_n$, so that the regression actually run is feasible, you still obtain

$$\sqrt{n}\left(\hat{\gamma}_{1,OLS} - \gamma_1\right) \xrightarrow{d} \mathcal{N}\left(0, \theta\right),$$

for some θ. Construct a consistent estimate of θ.

Hint: Find the joint asymptotic distribution of the OLS estimates $\left(\hat{\alpha}_0, \hat{\alpha}_1 - \hat{\alpha}_0, \hat{\beta}_0, \hat{\beta}_1 - \hat{\beta}_0\right)$ and $\bar{X}_n$, then use the continuous mapping theorem.

ANS: To ease notation let $\alpha_1 - \alpha_0 = \tau$ and let $\beta_1 - \beta_0 = \rho$. The difficulty with deriving a correct standard error comes from the fact that the estimation error in$\bar{X}_n$ does not disappear in the limiting distribution (essentially because it converges at the same rate as the OLS estimators). Our first task is to write $\hat{\gamma}_1 - \gamma_1$ in terms of the centered OLS estimates and $\bar{X}_n - \mathrm{E}(X)$. Note that

$$\begin{aligned}
\sqrt{n}\left(\hat{\gamma}_{1,OLS} - \gamma\right) &= \sqrt{n}\left(\hat{\tau} + \bar{X}_n'\hat{\rho} - \tau - \mathrm{E}(X)'\rho\right) \\
&= \sqrt{n}\left(\hat{\tau} - \tau\right) + \sqrt{n}\left(\bar{X}_n - \mathrm{E}(X)\right)'\hat{\rho} + \sqrt{n} \cdot \mathrm{E}(X)'\left(\hat{\rho} - \rho\right).
\end{aligned}$$

Now let $W_i = (1, D_i, X_i', D_i \cdot X_i')'$ and suppose that $X \in \mathbb{R}^k$. Note that despite changing W_i, ϵ_i remains unchanged because the two representations of the conditional mean of Y are equivalent -

all we did was reinterpret the parameters by altering $D \cdot X$ to $D \cdot (X - \mathrm{E}(X))$ in the final term.

$$\sqrt{n}\begin{pmatrix} \hat{\alpha}_0 - \alpha_0 \\ \hat{\tau} - \tau \\ \hat{\beta}_0 - \beta_0 \\ \hat{\rho} - \rho \\ \bar{X}_n - \mathrm{E}(X) \end{pmatrix} = \begin{pmatrix} \left[\frac{1}{n}\sum_{i=1}^n W_i W_i'\right]^{-1} & 0 \\ 0 & I_k \end{pmatrix} \frac{1}{\sqrt{n}} \sum_{i=1}^n \begin{pmatrix} W_i \epsilon_i \\ X_i - \mathrm{E}(X) \end{pmatrix}$$

$$\xrightarrow{d} \begin{pmatrix} \left[\mathrm{E}(W_i W_i')\right]^{-1} & 0 \\ 0 & I_k \end{pmatrix} \mathcal{N}(0_{3k+2}, V),$$

where

$$V = \begin{pmatrix} \left[\mathrm{E}(\epsilon_i^2 W_i W_i')\right] & 0 \\ 0 & Var(X_i) \end{pmatrix}.$$

The cross terms are zero because $\mathrm{E}\left(W\epsilon (X - \mathrm{E}(X))'\right) = \mathrm{E}\left(W(X - \mathrm{E}(X))' \mathrm{E}(\epsilon|D, X)\right) = 0$. Now we have to figure out how to represent $\sqrt{n}(\hat{\gamma}_{1,OLS} - \gamma_1)$ as a continuous function of the elements of the OLS estimates and $\bar{X}_n - \mathrm{E}(X)$. Observe that

$$\sqrt{n}(\hat{\gamma}_{1,OLS} - \gamma) = \left(0, 1, 0_k', \mathrm{E}(X)', \hat{\rho}'\right) \cdot \sqrt{n}\begin{pmatrix} \hat{\alpha}_0 - \alpha_0 \\ \hat{\tau} - \tau \\ \hat{\beta}_0 - \beta_0 \\ \hat{\rho} - \rho \\ \bar{X}_n - \mathrm{E}(X) \end{pmatrix} \xrightarrow{d} \mathcal{N}(0, \theta),$$

where

$$\theta = \left(0, 1, 0_k', \mathrm{E}(X)', \rho'\right) \begin{pmatrix} \left[\mathrm{E}(W_i W_i')\right]^{-1} \mathrm{E}(\epsilon_i^2 W_i W_i') \left[\mathrm{E}(W_i W_i')\right]^{-1} & 0 \\ 0 & Var(X_i) \end{pmatrix} \begin{pmatrix} 0 \\ 1 \\ 0_k \\ \mathrm{E}(X) \\ \rho \end{pmatrix}.$$

Note that the case $\rho = 0$ corresponds to the first case discussed in class where the slopes of the conditional means are zero, so a correct standard error could be obtained using the standard error for the coefficient on D in a regression of Y on a constant, D and X.

We can separate θ into a sum of two components. The first component represents the standard error we would obtain if we run the regression and specify robust. It's not large enough. Our corrected standard error is larger in this case because the second term is non-negative and the variance covariance matrix of the OLS estimates and $\bar{X}_n$ is block diagonal. Therefore, a consistent

```
.
. teffects ra (earn98 earn96 educ age married) (train), vce(robust)

Iteration 0:   EE criterion =  1.045e-27
Iteration 1:   EE criterion =  3.617e-31

Treatment-effects estimation                    Number of obs     =      1,130
Estimator      : regression adjustment
Outcome model  : linear
Treatment model: none

------------------------------------------------------------------------------
             |               Robust
      earn98 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
ATE          |
       train |
   (1 vs 0)  |   3.106202   .5862843     5.30   0.000     1.957106    4.255298
-------------+----------------------------------------------------------------
POmean       |
       train |
          0  |   9.179123   .2675471    34.31   0.000      8.65474    9.703506
------------------------------------------------------------------------------
```

Figure 1: Heteroskedasticity robust se computed using teffects

estimate of θ is given by

$$\hat{\theta} = \left(0, 1, 0_k', \bar{X}'\right) \left[\frac{W'W}{n}\right]^{-1} \left(\frac{1}{n}\sum_{i=1}^{n} \hat{\epsilon}_i^2 W_i W_i'\right) \left[\frac{W'W}{n}\right]^{-1} \begin{pmatrix} 0 \\ 1 \\ 0_k \\ \bar{X} \end{pmatrix} + \hat{\rho}' \widehat{Var}\left(X\right) \hat{\rho}.$$

The first part of this expression is the "HC0" standard error. Stata will report the HC1 standard error by default, which uses

$$\frac{1}{n-k}\sum_{i=1}^{n} \hat{\epsilon}_i^2 W_i W_i'$$

instead. Stata will also report the estimated variance covariance matrix of the regressors

$$\widehat{Var}\left(X\right) = \frac{1}{n-1}\sum_{i=1}^{n} \left(W_i - \bar{W}_n\right)\left(W_i - \bar{W}_n\right)'.$$

You could compute $\hat{\theta}$ using these reported variance matrices. Doing so would yield

$$\sqrt{\frac{\hat{\theta}}{n}} \approx 0.5902$$

Calculating HC0 standard errors for the first component of $\hat{\theta}$ instead would yield

$$\sqrt{\frac{\hat{\theta}}{n}} \approx 0.5876.$$

Stata has a treatment effects package (teffects) which will do this computation for you (Figure 1). The standard error reported is pretty similar to the one we computed using HC0 standard errors.

i) Use your previous answers to construct a test of $H_0 : \gamma_1 = 0$ vs. $H_1 : \gamma_1 \neq 0$ that is asymptotically of size α and justify your procedure formally.

ANS: By Slutsky's theorem,

$$\frac{\sqrt{n}\left(\hat{\gamma}_{1,OLS} - \gamma_1\right)}{\sqrt{\hat{\theta}}} \xrightarrow{d} \mathcal{N}\left(0,1\right).$$

Under H_0,

$$T_n = \frac{\sqrt{n}\hat{\gamma}_{1,OLS}}{\sqrt{\hat{\theta}}} \xrightarrow{d} \mathcal{N}\left(0,1\right),$$

so we reject iff $|T_n| > z_{1-\alpha/2}$.

j) Download the dataset jtrain98.dta from Canvas, and reproduce the regression results using the covariates used in class, this time allowing for the conditional means to have different slopes. Does this alteration drastically change the estimate of the ATE from the one presented in class? Is the result statistically significant at a 5% level?

Note: You will need to compute your consistent estimate of θ.

ANS: Yes, the estimated impact increases by about \$700 per year. Using the HC0 value of θ gives

$$|T_n| = \left|\frac{\sqrt{n}\hat{\gamma}_{1,OLS}}{\sqrt{\hat{\theta}}}\right| \approx 5.286 > z_{0.975},$$

which leads to a rejection of H_0 at the 5% level, so the result is statistically significant.

Question 3 Using the same setting as in Question 2, define $p(x) := \mathrm{P}(D = 1|X = x)$ and show that we can represent the ATE as

$$ATE = \mathrm{E}\left(\frac{Y\left(D - p\left(X\right)\right)}{p\left(X\right)\left(1 - p\left(X\right)\right)}\right).$$

Find similar representations for the ATT and ATU.

ANS: Looking at the form of the ATE suggests we consider $\mathrm{E}(YD)$. Since $p(X) = \mathrm{P}(D = 1|X)$, we should also try to condition on D and use the law of total expectation to get an expression that looks something like the formula given:

$$\begin{aligned}
\mathrm{E}\left(YD\right) &= \mathrm{E}\left[\mathrm{E}\left(YD|X\right)\right] \\
&= \mathrm{E}\left[\mathrm{E}\left(YD|D=1,X\right)\mathrm{P}\left(D=1|X\right) + \mathrm{E}\left(YD|D=0,X\right)\mathrm{P}\left(D=0|X\right)\right] \\
&= \mathrm{E}\left[\mathrm{E}\left(y_1|D=1,X\right)p\left(X\right)\right] \\
\text{(unconfoundedness)} &= \mathrm{E}\left[\mathrm{E}\left(y_1|X\right)p\left(X\right)\right].
\end{aligned}$$

It follows that

$$\mathrm{E}\left(\frac{YD}{p\left(X\right)}\right) = \mathrm{E}\left(y_1\right).$$

Similarly,

$$\mathrm{E}\left(\frac{Y\left(1-D\right)}{1-p\left(X\right)}\right) = \mathrm{E}\left(y_0\right).$$

Finally,

$$\frac{YD}{p(X)} - \frac{Y(1-D)}{1-p(X)} = \frac{Y(D-p(X))}{p(X)(1-p(X))}.$$

Note that we could have conditioned everything on X to begin with to obtain

$$\mathrm{E}(YD|X) = \mathrm{E}(y_1|X)\,p(X); \quad \mathrm{E}(Y(1-D)|X) = \mathrm{E}(y_0|X)[1-p(X)],$$

which yields

$$ATE(X) = \mathrm{E}(y_1 - y_0|X) = \mathrm{E}\left(\frac{Y(1-D)}{p(X)(1-p(X))}\middle| X\right).$$

To derive a similar expression for the ATT and ATU, note that

$$\begin{aligned}
ATT &= \mathrm{E}(ATE(X)|D=1) \\
&= \mathrm{E}\left[\mathrm{E}\left(\frac{Y(1-D)}{p(X)(1-p(X))}\middle| X\right)\middle| D=1\right] \\
&= \mathrm{E}\left[\mathrm{E}\left(\frac{Y(1-D)}{p(X)(1-p(X))}\middle| X\right)\frac{D}{\mathrm{P}(D=1)}\right] \\
&= \mathrm{E}\left[\mathrm{E}\left(\mathrm{E}\left(\frac{Y(1-D)}{p(X)(1-p(X))}\middle| X\right)\frac{D}{\mathrm{P}(D=1)}\middle| X\right)\right] \\
&= \frac{\mathrm{E}\left[\mathrm{E}\left(\frac{Y(1-D)}{p(X)(1-p(X))}\middle| X\right)\mathrm{P}(D=1|X)\right]}{\mathrm{E}(\mathrm{P}(D=1|X))}
\end{aligned}$$

Similarly,

$$\begin{aligned}
ATU &= \mathrm{E}\left[\mathrm{E}\left(\frac{Y(1-D)}{p(X)(1-p(X))}\middle| X\right)\frac{1-D}{\mathrm{P}(D=0)}\right] \\
&= \frac{\mathrm{E}\left[\mathrm{E}\left(\frac{Y(1-D)}{p(X)(1-p(X))}\middle| X\right)\mathrm{P}(D=0|X)\right]}{\mathrm{E}(\mathrm{P}(D=0|X))}.
\end{aligned}$$

We could stop at the third equality for the ATT to get an unconditional representation of the ATT, but the final representation of the ATT is used by Hirano, Imbens and Ridder (2003) to show that the ATT can be written as a weighted average treatment effect, where the weights are

$$\frac{\mathrm{P}(D=1|X)}{\mathrm{E}(\mathrm{P}(D=1|X))},$$

which integrate to 1.

Question 4 The dataset card.dta (available on canvas) contains data on wages, years of schooling and other observed characteristics of 2946 men with at least than 8 years of education. Consider the following model:

$$y = \beta_0 + \beta_1 x_1 + w'\gamma + u,$$

where $y = \ln(wage)$, x_1 = years of schooling and w is a set of included instruments that varies by specification. Estimate the following specifications using 2SLS and the optimal GMM estimator from

Lecture 3. In each specification the dependent variable is y, and the regressors include a constant, the endogenous variable x_1 and the following sets of included instruments. Report the coefficient estimate on x_1, as well as 95% confidence intervals for this coefficient, assuming homoskedasticity in the TSLS case and heteroskedasticity for the GMM estimates.

Specification 1: Included instruments: none; Excluded instruments: nearc4

Specification 2: Included instruments: south, smsa; Excluded instruments: nearc4

Specification 3: Included instruments: south, smsa; Excluded instruments: nearc4, nearc2

Specification 4: Included instruments: south, smsa, libcrd14, IQ, KWW, exper, expersq; Excluded instruments: nearc4, nearc2.

ANS: Dropping observations with values of `educ`< 8 leaves us with 2946 observations. In specification 4, the number of observations drops to 2034 after observations with missing values are omitted. The output from these regressions is:

Spec. 1	(1)	(2)
VARIABLES	TSLS Homosked	OGMM Hetero
lwage		
	-	-
educ	0.215***	0.215***
SE	(0.0342)	(0.0338)
95% CI	0.148 - 0.282	0.149 - 0.282
Constant	3.380***	3.380***
SE	(0.459)	(0.454)
95% CI	2.479 - 4.280	2.490 - 4.269
Observations	2,946	2,946

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Spec 2.	(1)	(2)
VARIABLES	TSLS Homosked	OGMM Hetero
lwage		
	-	-
educ	0.152**	0.152**
SE	(0.0715)	(0.0705)
95% CI	0.0122 - 0.292	0.0141 - 0.291
south	-0.101*	-0.101*
SE	(0.0566)	(0.0554)
95% CI	-0.212 - 0.0101	-0.209 - 0.00772
smsa	0.0596	0.0596
SE	(0.0589)	(0.0586)
95% CI	-0.0557 - 0.175	-0.0552 - 0.174
Constant	4.224***	4.224***
SE	(0.941)	(0.928)
95% CI	2.379 - 6.068	2.405 - 6.042
Observations	2,946	2,946

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Spec. 3	(1)	(2)
VARIABLES	TSLS Homosked	OGMM Hetero
lwage		
	-	-
educ	0.165**	0.164**
SE	(0.0736)	(0.0727)
95% CI	0.0206 - 0.309	0.0220 - 0.307
south	-0.0915	-0.0931
SE	(0.0583)	(0.0572)
95% CI	-0.206 - 0.0228	-0.205 - 0.0189
smsa	0.0500	0.0503
SE	(0.0606)	(0.0605)
95% CI	-0.0689 - 0.169	-0.0683 - 0.169
Constant	4.059***	4.064***
SE	(0.969)	(0.956)
95% CI	2.160 - 5.958	2.190 - 5.938
Observations	2,946	2,946

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Spec. 4	(1)	(2)
VARIABLES	TSLS Homosked	OGMM Hetero
lwage		
	-	-
educ	0.151**	0.144*
SE	(0.0756)	(0.0755)
95% CI	0.00294 - 0.299	-0.00409 - 0.292
south	-0.106***	-0.107***
SE	(0.0222)	(0.0221)
95% CI	-0.149 - -0.0625	-0.150 - -0.0636
smsa	0.128***	0.128***
SE	(0.0215)	(0.0212)
95% CI	0.0857 - 0.170	0.0866 - 0.170
libcrd14	-0.00510	-0.00434
SE	(0.0236)	(0.0228)
95% CI	-0.0513 - 0.0411	-0.0489 - 0.0403
IQ	0.000698	0.000773
SE	(0.00170)	(0.00169)
95% CI	-0.00263 - 0.00403	-0.00253 - 0.00408
KWW	-0.00331	-0.00248
SE	(0.00854)	(0.00858)
95% CI	-0.0201 - 0.0134	-0.0193 - 0.0143
exper	0.136***	0.133***
SE	(0.0446)	(0.0448)
95% CI	0.0489 - 0.224	0.0448 - 0.220
expersq	-0.00359***	-0.00354***
SE	(0.00108)	(0.00108)
95% CI	-0.00570 - -0.00149	-0.00566 - -0.00143
Constant	3.376***	3.466***
SE	(0.871)	(0.872)
95% CI	1.669 - 5.083	1.756 - 5.175
Observations	2,034	2,034
R-squared	0.113	0.130

Standard errors in parentheses
*** $p<0.01$, ** $p<0.05$, * $p<0.1$