1.

(a)

- Since $E(\bar{u}_i|\bar{x}_i) = E\left[E(\bar{u}_i|x_{i,1},\cdots,x_{i,n_i})\middle|\bar{x}_i\right]$

$$= E\left[\frac{1}{n_i}\sum_{e=1}^{n_i} E(u_{i,e}|x_{i,1},\cdots,x_{i,n_i})\middle|\bar{x}_i\right]$$

$$= E\left[\frac{1}{n_i}\times 0\middle|\bar{x}_i\right]$$

$$= 0,$$

$\hat{\beta}_1$ is unbiased.

For consistency,

$$\text{plim}(\hat{\beta}_1) = \beta_1 + \frac{Cov(\bar{x}_i, \bar{u}_i)}{Var(\bar{x}_i)}$$

$$Cov(\bar{u}_i, \bar{x}_i) = \frac{1}{n_i^2}\sum_{e=1}^{n_i}\sum_{f=1}^{n_i} Cov(u_{i,e}, x_{i,f})$$

$$= \frac{1}{n_i^2}\sum_{e=1}^{n_i}\sum_{f=1}^{n_i} Cov(v_{i,e}, x_{i,f}) = 0$$

So $\hat{\beta}_1$ is consistent.

- $Var(\bar{u}_i) = \sigma_f^2 + \frac{1}{n_i^2}\sum_{e=1}^{n_i} Var(u_{i,e})$ ($u_{i,e}$ are i.i.d.)

$$= \sigma_f^2 + \frac{1}{n_i^2}\cdot\sigma_v^2\cdot n_i$$

$$= \sigma_f^2 + \frac{\sigma_v^2}{n_i}$$

- $h_i = \sigma_f^2 + \frac{\sigma_v^2}{n_i} \Rightarrow \frac{\bar{y}_i}{\sqrt{n_i}} = \beta_0\frac{1}{\sqrt{n_i}} + \beta_1\frac{\bar{x}_i}{\sqrt{n_i}} + \frac{\bar{u}_i}{\sqrt{n_i}}$

$\hat{\beta}_1$ is more effcient.

(b)

- Still unbiased and consistent, because $(v_{i,e}, X_{i,e})$ still i.i.d. and f_i are still i.i.d.

- $Var(u_{i,e}) = Var(f_i + v_{i,e}) = Var(f_i) + Var(v_{i,e})$

$$= \sigma_f^2 + \sigma_v^2$$

$$\underset{i \neq j}{Cov}(u_{i,e}, u_{j,g}) = Cov(f_i, f_j) + Cov(f_i, v_{j,g}) + Cov(v_{i,e}, f_j) + Cov(v_{i,e}, v_{j,g})$$

$$= 0 + 0 + 0 + 0$$

$$= 0$$

$$Cov(u_{i,e}, u_{i,g}) = Var(f_i) + Cov(f_i, v_{i,g}) + Cov(v_{i,e}, f_i) + Cov(v_{i,e}, v_{i,g})$$

$$= \sigma_f^2 + 0 + 0 + 0$$

$$= \sigma_f^2$$

- Heteroskedasticity-robust standard errors.

- $Var(\hat{u}_{i,e}) = Var(u_{i,e} - \bar{u}_i)$

$$= Var\left(\frac{n_i - 1}{n_i} u_{i,e} - \frac{1}{n_i} \sum_{\substack{g \neq e \\ g=1}}^{n_i} u_{i,g}\right)$$

$$= \left(1 - \frac{1}{n_i}\right)^2 (\sigma_f^2 + \sigma_v^2) + \frac{1}{n_i^2} \cdot (\sigma_f^2 + \sigma_v^2)(n_i - 1) - 2\frac{n_i - 1}{n_i^2} \sigma_f^2 \cdot 1 - 2\frac{1}{n_i^2} \sigma_f^2 \cdot \frac{(n_i-1)(n_i-2)}{2}$$

$$= \left(1 - \frac{1}{n_i}\right) \sigma_v^2$$

$$Cov(\hat{u}_{i,e}, \hat{u}_{i,g}) = Cov(u_{i,e} - \bar{u}_i, u_{i,g} - \bar{u}_i) = Cov(u_{i,e}, u_{i,g}) - 2Cov(u_{i,e}, \bar{u}_i) + Var(\bar{u}_i)$$

$$= \sigma_f^2 - 2\left(\frac{1}{n_i}(\sigma_f^2 + \sigma_v^2) + \frac{n_i - 1}{n_i}\sigma_f^2\right) + \frac{1}{n_i^2} \cdot (\sigma_f^2 + \sigma_v^2) \cdot n_i + \frac{1}{n_i^2} \sigma_f^2 \cdot n_i(n_i - 1)$$

$$= 0 \cdot \sigma_f^2 - \frac{1}{n_i} \sigma_v^2$$

$$= -\frac{1}{n_i} \sigma_v^2$$

2.

(a) $E(y|x=1) = E(y|x=1, x^*=1) \times P(x^*=1|x=1) + E(y|x=1, x^*=0) \times P(x^*=0|x=1)$

$$= [\beta_0 + \beta_1 \times 1 + E(u|x=1, x^*=1)] \times P(x^*=1|x=1) +$$

$$[\beta_0 + \beta_1 \times 0 + E(u|x=1, x^*=0)] \times P(x^*=0|x=1)$$

$$= (\beta_0+\beta_1) P(x^*=1|x=1) + \beta_0 P(x^*=0|x=1)$$

$E(y|x=0) = E(y|x=0, x^*=1) \times P(x^*=1|x=0) + E(y|x=0, x^*=0) \times P(x^*=0|x=0)$

$$= (\beta_0+\beta_1) P(x^*=1|x=0) + \beta_0 P(x^*=0|x=0)$$

(b) $\hat{\beta_1} = \bar{y}_1 - \bar{y}_0 \xrightarrow{p} E(y|x=1) - E(y|x=0)$

$$= \beta_1 \times P(x^*=1|x=1) \quad \beta_1 \times P(x^*=1|x=0)$$

$$= \beta_1 \times [1 - P(x^*=0|x=1)] - \beta_1 \times [1 - P(x^*=0|x=0)]$$

$$= \beta_1 \times [P(x^*=0|x=0) - P(x^*=0|x=1)]$$

(c) $\hat{\beta_1} \xrightarrow{p} \beta_1$. Intution: no misclassification, $\hat{\beta_1}$ ectimator is unbiased.

(d) $\hat{\beta_1} \xrightarrow{p} -\beta_1$. Intuition: If the measurement is totally wrong, then it's actually "right", but reversed.

(e) $\hat{\beta_1} \xrightarrow{p} 0$. Intuition: The measurement has the same probility of 0 and 1, which is useless for OLS.

3.

(a) $y_i(1)$ is the outcome we would observe if i is trained

$y_i(0)$ - - - - - - - - - is not trained

$y_i(s)$ can be divided into 2 pares, $y_i(0)$ and $y_i(1)$

(b) Not statistically significant.

No.

(c) From the regression result, we know d is significantlly related to earn96, educ, age and married.

So d is not direct or only variable affecting y.

(d) $\hat{\tau}_{ATE} = 3.106202$, coefficient on d is 2.410547

4.

(c) Not much.

(d)

1) Yes.
2) Yes.
3) No.

5.

(c) se is halved.

(d) se returns to the same level in (b).

6.

$$\hat{\beta}^{IV} = \frac{Cov(z, y)}{Cov(z, x)} = \frac{\frac{1}{n}\sum_{i=1}^{n}(z_i - \bar{z})(y_i - \bar{y})}{\frac{1}{n}\sum_{i=1}^{n}(z_i - \bar{z})(x_i - \bar{x})}$$

$$= \frac{\sum z_i(y_i - \bar{y})}{\sum z_i(x_i - \bar{x})} = \frac{\sum z_i y_i - \bar{y}\sum z_i}{\sum z_i x_i - \bar{x}\sum z_i}$$

$$= \frac{n_1 \bar{y}_1 - \frac{n_0\bar{y}_0 + n_1\bar{y}_1}{n_0+n_1} n_1}{n_1 \bar{x}_1 - \frac{n_0\bar{x}_0 + n_1\bar{x}_1}{n_0+n_1} n_1} = \frac{(n_1^2 + n_0 n_1)\bar{y}_1 - n_0 n_1 \bar{y}_0 - n_1^2 \bar{y}_1}{(n_1^2 + n_0 n_1)\bar{x}_1 - n_0 n_1 \bar{x}_0 - n_1^2 \bar{x}_1}$$

$$= \frac{n_0 n_1}{n_0 n_1} \cdot \frac{\bar{y}_1 - \bar{y}_0}{\bar{x}_1 - \bar{x}_0}$$

$$= \frac{\bar{y}_1 - \bar{y}_0}{\bar{x}_1 - \bar{x}_0}$$

7.

(a) 0.0813949

(0.0041776)

(c) Yes. Since nearc4 is relevant for educ.

(d) 0.1703064

(0.0511423)

(e) IV is larger, with also larger se.