Solutions to Homework 3

1. (a) Note that

$$\mathrm{Cov}\left(X_i, u_i\right) = \mathrm{E}\left(X_i u_i\right) - \mathrm{E}\left(X_i\right)\mathrm{E}\left(u_i\right)$$

and given the condition $\mathrm{E}\left(u_i\right) = 0$, then $\mathrm{Cov}\left(X_i, u_i\right) \neq 0 \Leftrightarrow \mathrm{E}\left(X_i u_i\right) \neq 0$.

(b) Denote the IV estimator for β_0 and β_1 by $\hat{\beta}_{0,IV}$ and $\hat{\beta}_{1,IV}$ respectively, then using the condition that Z as the instrument variable for X should have $\mathrm{E}\left(Z_i u_i\right) = 0$. Therefore, moment conditions $\mathrm{E}\left(u_i\right)$ and $\mathrm{E}\left(X_i u_i\right) = 0$ imply the sample moment condition

$$\frac{1}{n}\sum\left(Y_i - \hat{\beta}_{0,IV} - \hat{\beta}_{1,IV} X_i\right) = 0, \quad \frac{1}{n}\sum Z_i\left(Y_i - \hat{\beta}_{0,IV} - \hat{\beta}_{1,IV} X_i\right) = 0,$$

which also suggests the corresponding normal system of equations

$$\begin{aligned} \sum Y_i &= n\hat{\beta}_{0,IV} + \hat{\beta}_{1,IV}\sum X_i \\ \sum Z_i Y_i &= \hat{\beta}_{0,IV}\sum Z_i + \hat{\beta}_{1,IV}\sum Z_i X_i. \end{aligned}$$

We can then solve $\hat{\beta}_{0,IV}$ and $\hat{\beta}_{1,IV}$ from the normal system of equations as

$$\hat{\beta}_{1,IV} = \frac{\sum z_i y_i}{\sum z_i x_i}, \quad \hat{\beta}_{0,IV} = \bar{Y} - \hat{\beta}_{1,IV}\bar{X}.$$

where $z_i = Z_i - \bar{Z}$ and $y_i = Y_i - \bar{Y}$.

(c) For this two step least square procedure, we have

$$\begin{aligned} \tilde{\beta}_1 &= \frac{\sum \hat{x}_i y_i}{\sum \hat{x}_i^2} \\ &= \frac{\sum\left(\hat{\alpha}_0 + \hat{\alpha}_1 Z_i - \bar{\hat{X}}\right) y_i}{\sum\left(\hat{\alpha}_0 + \hat{\alpha}_1 Z_i - \bar{\hat{X}}\right)\left(X_i - e_i - \bar{X}\right)} \\ &= \frac{\sum\left(\hat{\alpha}_0 + \hat{\alpha}_1 Z_i - \hat{\alpha}_0 - \hat{\alpha}_1\bar{Z}\right) y_i}{\sum\left(\hat{\alpha}_0 + \hat{\alpha}_1 Z_i - \hat{\alpha}_0 - \hat{\alpha}_1\bar{Z}\right)\left(X_i - e_i - \bar{X}\right)} \\ &= \frac{\sum \hat{\alpha}_1 z_i y_i}{\sum \hat{\alpha}_1 z_i x_i - \sum \hat{\alpha}_1 z_i e_i} = \frac{\sum z_i y_i}{\sum z_i x_i} = \hat{\beta}_{1,IV} \end{aligned}$$

where $\bar{\hat{X}} = \frac{1}{n}\sum \hat{X}_i$, $\hat{x}_i = \hat{X}_i - \bar{\hat{X}}$, $x_i = X_i - \bar{X}$, $y_i = Y_i - \bar{Y}$, and $z_i = Z_i - \bar{Z}$. Besides, e_i refers to the corresponding residual when regressing X on Z and hence $\sum z_i e_i = 0$. Given the expression for $\tilde{\beta}_1$, it is possible to solve $\tilde{\beta}_0$ as

$$\tilde{\beta}_0 = \bar{Y} - \tilde{\beta}_1 \bar{\hat{X}} = \bar{Y} - \hat{\beta}_{1,IV}\bar{X} = \hat{\beta}_{0,IV},$$

where we have used the fact that from the first step least square routine $\frac{1}{n}\sum X_i = \frac{1}{n}\sum \hat{X}_i$.

2. (a) For the averaged data, we can run the following regression model

$$\bar{Y}_g = \beta_0 + \beta_1 \bar{X}_g + \bar{u}_g$$

where $\bar{Y}_g$ signifies the average of Y's within the gth village, and $\bar{X}_g$ and $\bar{u}_g$ are similarly defined. Under the classical assumptions, u's are serially correlated and homoskedastic with variance σ^2, implying that $\bar{u}_g$ are heteroskedastic with variance σ^2/n_g. As a consequence, the conventional standard error is inconsistent and inference based on it will be invalid and misleading.

(b) Multiplying $\sqrt{n_g}$ on both sides of $\bar{Y}_g = \beta_0 + \beta_1 \bar{X}_g + \bar{u}_g$ yields

$$\tilde{Y}_g = \beta_0 + \beta_1 \tilde{X}_g + \tilde{u}_g$$

where $\tilde{Y}_g = \sqrt{n_g}\bar{Y}_g$, $\tilde{X}_g = \sqrt{n_g}\bar{X}_g$, $\tilde{u}_g = \sqrt{n_g}\bar{u}_g$. Then $\operatorname{Var}(\tilde{u}_g) = n_g \operatorname{Var}(\bar{u}_g) = \sigma^2$ implying that $\tilde{u}_g$ is homoskedastic and the corresponding OLS estimator in $\tilde{Y}_g = \beta_0 + \beta_1 \tilde{X}_g + \tilde{u}_g$ is BLUE and therefore more efficient than the OLS estimator in $\bar{Y}_g = \beta_0 + \beta_1 \bar{X}_g + \bar{u}_g$.

3. (a) Since now the error term is now given by $v_i = u_i + e_i$. $\{u_i\}$ is IID, $\mathrm{E}(u_i \mid X_i) = 0$ and $\operatorname{Var}(u_i \mid X_i) = \sigma_u^2$. e_i is independent of u_i and X_i, we have

$$\mathrm{E}(v_i \mid X_i) = \mathrm{E}(e_i + u_i \mid X_i) = 0$$

and

$$\sigma_v^2 \equiv \operatorname{Var}(v_i \mid X_i) = \operatorname{Var}(e_i \mid X_i) + \operatorname{Var}(u_i \mid X_i) = \sigma_e^2 + \sigma_u^2.$$

Let $\boldsymbol{\beta} = (\beta_0, \beta_1)'$, the OLS estimator based on $\left\{\tilde{Y}_i\right\}$ is given by

$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}'\boldsymbol{X}\right)' \boldsymbol{X}'\tilde{\boldsymbol{Y}} = \boldsymbol{\beta} + \boldsymbol{X}'\boldsymbol{v}$$

where we use $\boldsymbol{v}$ to denote $\{v_i\}$. Given the structure of v_i, we have $\mathrm{E}(\boldsymbol{X}'\boldsymbol{v} \mid \boldsymbol{X}) = 0$ and $(1/n)\boldsymbol{X}'\boldsymbol{v} \xrightarrow{p} 0$, therefore $\hat{\boldsymbol{\beta}}$ is unbiased and consistent.

(b) Yes. Note that we do not have conditional heteroskedasticity and we have IID observations here. So we can construct the confidence interval as usual — there is no need to consider the White estimator for the s.e. of the OLS estimator.

(c) This is a true statement. Measurement error in the regressor X_i leads to inconsistent results in general. Measurement error in Y_i does not affect consistency, as long as the measurement error is not correlated with the regressors.

(d) Because v_i has a greater (conditional) variance than u_i, the standard errors of the OLS estimator $\hat{\boldsymbol{\beta}}$ would be bigger than the case without measurement error in the dependent variable. The confidence intervals would be wider as a result. The coefficient estimates should not change too much because the estimators are consistent. In addition, the R^2 will be smaller, reflecting the larger variance of the regression errors.