AS.180.633: Econometrics **Spring 2020**

Homework 4: Suggested Solutions

Instructor: Yingyao Hu *By:* Tong Zhou

4.16

Take the linear homoskedastic CEF

$$y_i^* = \boldsymbol{x}_i'\boldsymbol{\beta} + e_i \tag{1}$$
$$\mathbb{E}[e_i|\boldsymbol{x}_i] = 0$$
$$\mathbb{E}[e_i^2|\boldsymbol{x}_i] = \sigma^2$$

and suppose that y_i^* is measured with error. Instead of y_i^*, we observe y_i which satisfies

$$y_i = y_i^* + u_i$$

where u_i is measurement error. Suppose that e_i and u_i are independent and

$$\mathbb{E}[u_i|\boldsymbol{x}_i] = 0;$$
$$\mathbb{E}[u_i^2|\boldsymbol{x}_i] = \sigma_u^2(\boldsymbol{x}_i)$$

1. Derive an equation for y_i as a function of $\boldsymbol{x}_i$. Be explicit to write the error term as a function of the structural errors e_i and u_i. What is the effect of this measurement error on the model Eq. (1)
2. Describe the effect of this measurement error on OLS estimation of $\boldsymbol{\beta}$ in the feasible regression of the observed y_i on $\boldsymbol{x}_i$.
3. Describe the effect (if any) of this measurement error on standard error calculation for $\widehat{\boldsymbol{\beta}}$.

Proof. 1. Define a composite error

$$v_i = e_i + u_i,$$

then

$$y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + v_i$$

and the assumption $\mathbb{E}[v_i|\boldsymbol{x}_i] = 0$ is still maintained.

2. Regress y on x, the estimator is obtained:

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$$

From $\mathbb{E}[v_i|x_i] = 0$,

$$\mathbb{E}\left[\widehat{\boldsymbol{\beta}}\middle|\boldsymbol{X}\right] = \boldsymbol{\beta}.$$

So the estimator is unbiased for $\boldsymbol{\beta}$ despite the presence of measurement error in $\boldsymbol{y}$.

3. The variance of $\widehat{\boldsymbol{\beta}}$ is:

$$\begin{aligned}\mathsf{Var}\left(\widehat{\boldsymbol{\beta}}\middle|\boldsymbol{X}\right) &= \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\underbrace{\mathbb{E}[vv'|\boldsymbol{X}]}_{=\sigma^2\boldsymbol{I}_n+\boldsymbol{\Omega}_u}\boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \\ &= \sigma^2\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} + \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{\Omega}_u\boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\end{aligned}$$

where $\boldsymbol{\Omega}_u = \mathrm{diag}(\sigma_u^2(x_1), \ldots, \sigma_u^2(x_n))$.

Thus, the presence of measurement error in y does not affect unbiasedness but inflates variances. □

4.20

Take the model

$$\begin{aligned}\boldsymbol{y} &= \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e} \\ \mathbb{E}[\boldsymbol{e}|\boldsymbol{X}] &= \boldsymbol{0} \\ \mathbb{E}[\boldsymbol{e}\boldsymbol{e}'|\boldsymbol{X}] &= \boldsymbol{\Omega}.\end{aligned}$$

Assume for simplicity that $\boldsymbol{\Omega}$ is known. Consider the OLS and GLS estimators $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}(\boldsymbol{X}'\boldsymbol{y})$ and $\widetilde{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{X})^{-1}(\boldsymbol{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{y})$. Compute the (conditional) covariance between $\widehat{\boldsymbol{\beta}}$ and $\widetilde{\boldsymbol{\beta}}$:

$$\mathbb{E}[(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta})'|\boldsymbol{X}]$$

Find the (conditional) covariance matrix for $\widehat{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}}$:

$$\mathbb{E}[(\widehat{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}})(\widehat{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}})'|\boldsymbol{X}].$$

Proof. By

$$\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{e} \tag{2}$$

$$\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta} = (\boldsymbol{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{e}, \tag{3}$$

we have

$$\begin{aligned}
\mathbb{E}[(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta})'|\boldsymbol{X}] &= \mathbb{E}[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{e}\boldsymbol{e}'\boldsymbol{\Omega}^{-1}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{X})^{-1}|\boldsymbol{X}] \\
&= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' \underbrace{\mathbb{E}[\boldsymbol{e}\boldsymbol{e}'|\boldsymbol{X}]}_{=\boldsymbol{\Omega}} \boldsymbol{\Omega}^{-1}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{X})^{-1} \\
&= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{X})^{-1} \\
&= (\boldsymbol{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{X})^{-1}
\end{aligned}$$

For the second part,

$$\begin{aligned}
\mathbb{E}[(\widehat{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}})(\widehat{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}})'|\boldsymbol{X}] &= \mathbb{E}[(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} + \boldsymbol{\beta} - \widetilde{\boldsymbol{\beta}})(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} + \boldsymbol{\beta} - \widetilde{\boldsymbol{\beta}})'|\boldsymbol{X}]. \\
&= \underbrace{\mathbb{E}[(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})'|\boldsymbol{X}]}_{=\mathrm{Var}(\widehat{\boldsymbol{\beta}}|\boldsymbol{X})} - \mathbb{E}[(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta})'|\boldsymbol{X}] \\
&\quad - \mathbb{E}[(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta})(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})'|\boldsymbol{X}] + \underbrace{\mathbb{E}[(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta})(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta})'|\boldsymbol{X}]}_{=\mathrm{Var}(\widetilde{\boldsymbol{\beta}}|\boldsymbol{X})} \\
&= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Omega}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} - (\boldsymbol{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{X})^{-1} \\
&\quad - (\boldsymbol{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{X})^{-1} + (\boldsymbol{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{X})^{-1} \\
&= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Omega}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} - (\boldsymbol{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{X})^{-1}
\end{aligned}$$

□

4.23

Take the linear regression model with $\mathbb{E}[\boldsymbol{y}|\boldsymbol{X}] = \boldsymbol{X}\boldsymbol{\beta}$. Define the *ridge regression* estimator

$$\widehat{\boldsymbol{\beta}} = \left(\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{I}_k\lambda\right)^{-1}\boldsymbol{X}'\boldsymbol{y}$$

where $\lambda > 0$ is a fixed constant. Find $\mathbb{E}[\widehat{\boldsymbol{\beta}}\,|\,\boldsymbol{X}]$. Is $\widehat{\boldsymbol{\beta}}$ biased for $\boldsymbol{\beta}$?

Proof. Since

$$\mathbb{E}[\widehat{\boldsymbol{\beta}}\,|\,\boldsymbol{X}] = \left(\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{I}_k\lambda\right)^{-1}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} \neq \boldsymbol{\beta},$$

it is therefore biased for $\boldsymbol{\beta}$. □

4.26

Extend the empirical analysis reported in Section 4.22. Do a regression of standardized test score (*totalscore* normalized to have zero mean and variance 1) on tracking, age, gender, being assigned to the contract teacher, and student's percentile in the initial distribution. (The sample size will be smaller as some observation have missing variables.) Calculate standard errors using both the conventional robust formula, and clustering based on the school.

1. Compare the two sets of standard errors. Which standard error changes the most by clustering? Which changes the least?
2. How does the coefficient on *tracking* change by inclusion of the individual controls (in comparison to the results from (4.54))?

See Table 1 for comparison of the clustered s.e. and the conventional HC1 s.e., from which *tracking* changes the most and *gender* changes the least. As a consequence, the p-value of the estimate for *tracking* was 0.026 when the clustered standard errors were used, implying that it was insignificantly different from 0 under 1% significance level.

At the same time, the estimate for *tracking* was 0.174, while it was 0.138 in (4.54). More striking was the change in p-value.

Table 1: Comparison of clustered s.e. and HC1 s.e.

	Coefficients	Clustered s.e.	HC1 s.e.	change by Δ%
Intercept	-0.742	0.131	0.082	67.8%
tracking	0.174	0.077	0.024	226.5%
assigned	0.181	0.038	0.024	65.5%
age	-0.041	0.013	0.009	23.5%
gender	0.081	0.029	0.024	18.9%
percentile	0.017	0.0007	0.0004	177.6%

(See R codes on the next page)

R codes:

```
library(scales)
library(stargazer)

ddk <- read.csv("DDK2011.csv",header= TRUE, sep=",")

## For non-numeric variables, do not forget as.character(),
   otherwise as.numeric() would pump up wrong numbers.

testscore <- as.matrix(ddk$totalscore)
tracking <- as.numeric(as.character(ddk$tracking))
etpteacher <- as.numeric(as.character(ddk$etpteacher))
agetest <- as.numeric(as.character(ddk$agetest))
girl <- as.numeric(as.character(ddk$girl))
percentile <- as.numeric(as.character(ddk$percentile))
schoolid <- as.matrix(ddk$schoolid)

ddk2 <- na.omit(cbind(matrix(1,n,1),tracking,etpteacher,agetest,
   girl,percentile,testscore,schoolid))

y <- scale(ddk2[,7])
n <- nrow(y)

x <- ddk2[,1:6]

schoolid <- ddk2[,8]

k <- ncol(x)

xx <- t(x)%*%x

invx <- solve(xx)

beta <- solve(xx,t(x)%*%y)

## Clustered s.e.

xe <- x*rep(y-x%*%beta,times=k)
xe_sum <- rowsum(xe,schoolid)
G <- nrow(xe_sum)
omega <- t(xe_sum)%*%xe_sum
scale <- G/(G-1)*(n-1)/(n-k)
V_clustered <- scale*invx%*%omega%*%invx
```

```
se_clustered <- sqrt(diag(V_clustered))

## Conventional HC1 s.e.

e = y-x%*%beta;
a <- n/(n-k)
u1 <- x*(e%*%matrix(1,1,k))

v1a <- a * invx %*% (t(u1)%*%u1) %*% invx
s1a <- sqrt(diag(v1a))

change_by <- percent((se_clustered - s1a)/s1a)

compare <- cbind(beta,se_clustered, s1a,change_by)

colnames(compare) <- c("Coefficients","Clustered s.e.", "HC1 s.e.","
    change by")

rownames(compare) <- c("Intercept","tracking","assigned","age","
    gender","percentile")

stargazer(compare)
```