Solutions for Homework 4

Tong Zhou

Problems: 4.16, 4.18, 4.23, 4.25, 4.26, 5.5.

PROBLEM 4.16

Proof. 1. Define a composite error

$$\nu_i = e_i + u_i,$$

then

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

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

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

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

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

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

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

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

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

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

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

PROBLEM 4.18

Proof. Define the short regression as

$$\boldsymbol{y} = \boldsymbol{X}_1\beta_1 + \boldsymbol{e}_1,$$

where $\boldsymbol{y} = \begin{pmatrix} y_1 & y_2 & \cdots & y_n \end{pmatrix}'$, $\boldsymbol{X}_1 = \begin{pmatrix} X_1 & \cdots & X_n \end{pmatrix}'$, $\boldsymbol{e}_1 = \begin{pmatrix} e_{11} & \cdots & e_{1n} \end{pmatrix}'$. Here the dimensions $\boldsymbol{X}_1$ and $\boldsymbol{e}_1$ are $n \times k_1$ and $n \times 1$ respectively. The same notations apply to $\boldsymbol{X}_2$ and $\boldsymbol{y}$.

Its residual

$$\widehat{\boldsymbol{e}}_1 = \boldsymbol{M}_1 \boldsymbol{e}_1,$$

where $\boldsymbol{M}_1 = I_n - \boldsymbol{X}_1(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'$ and $\boldsymbol{e}_1 = \boldsymbol{X}_2\beta_2 + \boldsymbol{e}$.

Thus we have

$$\mathbb{E}[s^2 \mid \boldsymbol{X}] = \frac{1}{n-k_1}\mathbb{E}[\boldsymbol{e}_1'\boldsymbol{M}_1\boldsymbol{e}_1 \mid \boldsymbol{X}].$$

Now it suffices to look into the term $\mathbb{E}[\boldsymbol{e}_1'\boldsymbol{M}_1\boldsymbol{e}_1 \mid \boldsymbol{X}]$:

$$\begin{aligned}
\mathbb{E}[\boldsymbol{e}_1'\boldsymbol{M}_1\boldsymbol{e}_1 \mid \boldsymbol{X}] &= \mathbb{E}[(\boldsymbol{X}_2\beta_2 + \boldsymbol{e})'\boldsymbol{M}_1(\boldsymbol{X}_2\beta_2 + \boldsymbol{e}) \mid \boldsymbol{X}] \\
&= \mathbb{E}[\,\beta_2'\boldsymbol{X}_2'\boldsymbol{M}_1\boldsymbol{X}_2\beta_2 + \boldsymbol{e}'\boldsymbol{M}_1\boldsymbol{e} + 2\beta_2'\boldsymbol{X}_2'\boldsymbol{M}_1\boldsymbol{e} \mid \boldsymbol{X}] \\
&= \beta_2'\boldsymbol{X}_2'\boldsymbol{M}_1\boldsymbol{X}_2\beta_2 + \mathbb{E}[\boldsymbol{e}'\boldsymbol{M}_1\boldsymbol{e} \mid \boldsymbol{X}] + 2\beta_2'\boldsymbol{X}_2'\boldsymbol{M}_1\underbrace{\mathbb{E}[\boldsymbol{e} \mid \boldsymbol{X}]}_{=\boldsymbol{0}} \\
&= \beta_2'\boldsymbol{X}_2'\boldsymbol{M}_1\boldsymbol{X}_2\beta_2 + (n-k_1)\sigma^2.
\end{aligned}$$

As a result, we have

$$\mathbb{E}[s^2 \mid \boldsymbol{X}] = \sigma^2 + \frac{1}{n-k_1}\beta_2'\boldsymbol{X}_2'\boldsymbol{M}_1\boldsymbol{X}_2\beta_2. \tag{1}$$

Equation (1) implies that s^2 in general is an upward biased estimator for σ^2, since $\beta_2'\boldsymbol{X}_2'\boldsymbol{M}_1\boldsymbol{X}_2\beta_2$ is non-negative. □

PROBLEM 4.23

PROOF. Since

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

it is therefore biased for β. □

PROBLEM 4.25

See below.

PROBLEM 4.26

See below.

PROBLEM 5.5

PROOF. From Hansen's Theorem 5.4,

$$\widehat{\beta} \sim \mathcal{N}(\beta, \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}).$$

Then conditional on $\boldsymbol{X}$, $\boldsymbol{x}_i'$ becomes constant for each i. By theorem 5.2, it follows that $\widehat{y}_i = \boldsymbol{x}_i'\widehat{\beta}$ follows a normal distribution such that

$$\widehat{y}_i \mid \boldsymbol{X} \sim \mathcal{N}(\boldsymbol{x}_i'\beta, \sigma^2 h_{ii}).$$

□

Codes for Problems 4.25 and 4.26

```
rm(list=ls())
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --
## v ggplot2 3.3.3     v purrr   0.3.4
## v tibble  3.1.0     v dplyr   1.0.4
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(knitr)
library(gtsummary)   ## Summarize regression results
library(moderndive)  ## Summarize regression results
library(haven)       ## Use function read_stata to load dataset.
library(jtools)      ## Report robust standard errors.
library(huxtable)
```

```
##
## Attaching package: 'huxtable'
## The following object is masked from 'package:gtsummary':
##
##     as_flextable
## The following object is masked from 'package:dplyr':
##
##     add_rownames
## The following object is masked from 'package:ggplot2':
##
##     theme_grey
```

Problem 4.25

```
df <- read_stata("~/Dropbox/2021_Spring/Econometrics/hw3/cps09mar.dta")
#### create new variables: log_wage, experience, experience2
df <- df %>% mutate(log_wage = log(earnings/(hours*week)),
                    exper = age - education - 6,
                    exper2 = exper^2/100)



#### create new dummy variables
df <- df %>%
  mutate(d_NE = ifelse(region==1,1,0),
```

```
        d_S = ifelse(region==3,1,0),
        d_W = ifelse(region==4,1,0),
        d_married = ifelse(marital==1 | marital==2,1,0),
        d_widow= ifelse(marital==4 | marital==5,1,0),
        d_sep = ifelse(marital==6,1,0)) %>%
  select(-age)
```

Report the HC3 standard error:

```
#### do regression and obtain table
df %>% filter(race==1,female==0,hisp==1) %>%
  lm(log_wage~education+exper+exper2+d_NE+d_S+
       d_W+d_married+d_widow+d_sep, data=.)  %>%
  export_summs(.robust="HC3",digits=3,model.names="Problem 4.25, HC3")
```

Problem 4.26

Load data and standardize the score variable:

```
ddk <- read_stata("~/Dropbox/2021_Spring/Econometrics/hw4/ddk2011.dta")

ddk <- ddk %>% mutate(std_score = scale(totalscore))
```

Tow regressions: conventional s.e. and clustered s.e.

```
model1 <-  ddk %>% lm(std_score~tracking+etpteacher+agetest+girl+percentile,data=.)

model2 <- ddk %>% lm(std_score~tracking+etpteacher+agetest+girl+percentile,data=.)
```

Report two regression results

```
 export_summs(model1,model2,robust=c("HC3","HC3"),cluster=c("NULL","schoolid"),
            digits=3,model.names = c("Conventional s.e.","Clustered s.e."))
```

	Problem 4.25, HC3
(Intercept)	1.208 ***
	(0.051)
education	0.087 ***
	(0.003)
exper	0.028 ***
	(0.003)
exper2	-0.036 ***
	(0.005)
d_NE	0.063
	(0.038)
d_S	-0.066 *
	(0.031)
d_W	0.018
	(0.030)
d_married	0.191 ***
	(0.022)
d_widow	0.091 *
	(0.041)
d_sep	0.020
	(0.058)
N	4230
R2	0.252

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$.

	Conventional s.e.	Clustered s.e.
(Intercept)	-0.729 ***	-0.729 ***
	(0.081)	(0.132)
tracking	0.173 ***	0.173 *
	(0.024)	(0.077)
etpteacher	0.180 ***	0.180 ***
	(0.024)	(0.038)
agetest	-0.041 ***	-0.041 **
	(0.009)	(0.014)
girl	0.081 ***	0.081 **
	(0.024)	(0.029)
percentile	0.017 ***	0.017 ***
	(0.000)	(0.001)
N	5269	5269
R2	0.249	0.249

Standard errors are heteroskedasticity robust. *** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$.