Empirical Exercise - E7.1

Chi-Yuan Fang

2021-03-28

Use the **Birthweight_Smoking** data set introduced in Empirical Exercise E5.3 to answer the following questions. To begin, run three regressions:

(1) *Birthweight* on *Smoker*
(2) *Birthweight* on *Smoker*, *Alcohol*, and *Nprevist*
(3) *Birthweight* on *Smoker*, *Alcohol*, *Nprevist*, and *Unmarried*

a. What is the value of the estimated effect of smoking on birth weight in each of the regressions?

Solution

```
library(kableExtra); library(stargazer)
library(ggplot2); library(dplyr);
library(jtools); library(ggstance);
library(broom.mixed); library(huxtable)
```

```
# import data set
library(readxl)
Birthweight_Smoking <- read_excel("Birthweight_Smoking/Birthweight_Smoking.xlsx")

# variable names
colnames(Birthweight_Smoking)
```

```
##  [1] "nprevist"    "alcohol"     "tripre1"     "tripre2"     "tripre3"
##  [6] "tripre0"     "birthweight" "smoker"      "unmarried"   "educ"
## [11] "age"         "drinks"
```

```
# https://cran.r-project.org/web/packages/stargazer/stargazer.pdf
# https://www.jakeruss.com/cheatsheets/stargazer/
# https://cran.r-project.org/web/packages/stargazer/vignettes/stargazer.pdf

library(sandwich); library(lmtest)
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
# Model 1
E71_M1 <- lm(formula = birthweight ~ smoker,
             data = Birthweight_Smoking)

# Model 2
E71_M2 <- lm(formula = birthweight ~ smoker + alcohol + nprevist,
```

```
            data = Birthweight_Smoking)

# Model 3
E71_M3 <- lm(formula = birthweight ~ smoker + alcohol + nprevist + unmarried,
             data = Birthweight_Smoking)

# Adjust standard errors
cov1 <- vcovHC(E71_M1, type = "HC1")
rb_se1 <- sqrt(diag(cov1))

cov2 <- vcovHC(E71_M2, type = "HC1")
rb_se2 <- sqrt(diag(cov2))

cov3 <- vcovHC(E71_M3, type = "HC1")
rb_se3 <- sqrt(diag(cov3))

# output table
stargazer(E71_M1, E71_M2, E71_M3,
          type = "text",
          se = list(rb_se1, rb_se2, rb_se3),
          digits = 4)
```

```
##
## ================================================================================================
##                                            Dependent variable:
##                     ---------------------------------------------------------------------------
##                                                birthweight
##                                (1)                    (2)                    (3)
## ------------------------------------------------------------------------------------------------
## smoker                     -253.2284***           -217.5801***           -175.3769***
##                              (26.8104)              (26.1076)              (26.8268)
##
## alcohol                                             -30.4913               -21.0835
##                                                     (72.5967)              (72.9921)
##
## nprevist                                            34.0699***             29.6025***
##                                                     (3.6083)               (3.5827)
##
## unmarried                                                                  -187.1332***
##                                                                             (27.6772)
##
## Constant                   3,432.0600***          3,051.2490***          3,134.4000***
##                              (11.8905)              (43.7145)              (44.1486)
##
## ------------------------------------------------------------------------------------------------
## Observations                  3,000                  3,000                  3,000
## R2                            0.0286                 0.0729                 0.0886
## Adjusted R2                   0.0283                 0.0719                 0.0874
## Residual Std. Error    583.7297 (df = 2998)   570.4708 (df = 2996)   565.6975 (df = 2995)
## F Statistic         88.2793*** (df = 1; 2998) 78.4697*** (df = 3; 2996) 72.7930*** (df = 4; 2995)
## ================================================================================================
## Note:                                                              *p<0.1; **p<0.05; ***p<0.01
```

b. Construct a 95% confidence interval for the effect of smoking on birth weight, using each of

the regressions.

Solution

```
# https://stats.stackexchange.com/questions/117052/replicating-statas-robust-option-in-r
library(estimatr)

# Model 1 # use HC1
E71_M1n <- lm_robust(formula = birthweight ~ smoker,
                     data = Birthweight_Smoking,
                     se_type = "stata")

# Model 2
E71_M2n <- lm_robust(formula = birthweight ~ smoker + alcohol + nprevist,
                     data = Birthweight_Smoking,
                     se_type = "stata")

# Model 3
E71_M3n <- lm_robust(formula = birthweight ~ smoker + alcohol + nprevist + unmarried,
                     data = Birthweight_Smoking,
                     se_type = "stata")

export_summs(E71_M1n, E71_M2n, E71_M3n,
             digits = 4,
             ci_level = 0.95,
             error_format = "[{conf.low}, {conf.high}]",
             model.names = c("Model (1)", "Model (2)", "Model (3)"))
```

c. Does the coefficient on Smoker in regression (1) suffer from omitted variable bias? Explain.

Solution

```
(E71_M2$coefficients[2] - E71_M1$coefficients[2]) / E71_M1$coefficients[2]
```

```
##     smoker
## -0.1407752
```

Yes, it does. The coefficient changes by roughly 14% in magnitude when additional regressors are added to (1). This change is substantively large and large relative to the standard error in (1).

d. Does the coefficient on Smoker in regression (2) suffer from omitted variable bias? Explain.

Solution

```
(E71_M3$coefficients[2] - E71_M2$coefficients[2]) / E71_M2$coefficients[2]
```

```
##     smoker
## -0.1939661
```

Yes, it does. The coefficient changes by roughly 19% in magnitude when *unmarried* is added as an additional regression. This change is substantively large and large relative to the standard error in (2).

e. Consider the coefficient on *Unmarried* in regression (3).
   i. Construct a 95% confidence interval for the coefficient.
   ii. Is the coefficient statistically significant? Explain.
   iii. Is the magnitude of the coefficient large? Explain.
   iv. A family advocacy group notes that the large coefficient suggests that public policies that encourage marriage will lead, on average, to healthier babies. Do you agree? (Hint: Review the discussion of control variables in Section 6.8. Discuss some of the various

	Model (1)	Model (2)	Model (3)
(Intercept)	3432.0600 ***	3051.2486 ***	3134.4000 ***
	[3408.7455, 3455.3744]	[2965.5352, 3136.9619]	[3047.8354, 3220.9646]
smoker	-253.2284 ***	-217.5801 ***	-175.3769 ***
	[-305.7970, -200.6597]	[-268.7708, -166.3894]	[-227.9777, -122.7761]
alcohol		-30.4913	-21.0835
		[-172.8357, 111.8531]	[-164.2032, 122.0363]
nprevist		34.0699 ***	29.6025 ***
		[26.9949, 41.1450]	[22.5777, 36.6274]
unmarried			-187.1332 ***
			[-241.4014, -132.8651]
nobs	3000	3000	3000
r.squared	0.0286	0.0729	0.0886
adj.r.squared	0.0283	0.0719	0.0874
statistic	89.2110	59.4841	56.0850
p.value	0.0000	0.0000	0.0000
df.residual	2998.0000	2996.0000	2995.0000
nobs.1	3000.0000	3000.0000	3000.0000
se_type	HC1.0000	HC1.0000	HC1.0000

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

factors that *Unmarried* may be controlling for and how this affects the interpretation of its coefficient.)

Solution

i. See part (b).

ii.

- **Prepare**

  $H_0 : \beta_{unmarried} = 0$ v.s. $H_1 : \beta_{unmarried} \neq 0$

  Let the significance level be 5%.

- **Calculate**

  See the results in part (a) or (b).

- **Conclusion**

Because $p - value < \alpha = 5\%$, we reject H_0. There is significant evidence that the coefficient on *unmarried* is different from 0.

iii. Yes, it is. Other things being equal, *birthweight* is 187 grams lower for unmarried mothers on average.

iv. As the question suggests, *unmarried* is a control variable that captures the effects of several factors that differ between married and unmarried mothers such as age, education, income, diet and other health factors, and so forth.

f. Consider the various other control variables in the data set. Which do you think should be included in the regression? Using a table like Table 7.1, examine the robustness of the confidence interval you constructed in (b). What is a reasonable 95% confidence interval for the effect of smoking on birth weight?

Solution

We consider adding on additional regression in the table that includes *Age* and *Educ* (years of education) in model (4).

```
# Model 4
E71_M4 <- lm(formula = birthweight ~ smoker + alcohol + nprevist + unmarried + age + educ,
             data = Birthweight_Smoking)

# Adjust standard errors
cov4 <- vcovHC(E71_M4, type = "HC1")
rb_se4 <- sqrt(diag(cov4))

# output table
stargazer(E71_M3, E71_M4,
          type = "text",
          column.labels=c("Model (3)","Model (4)"),
          se = list(rb_se3, rb_se4),
          digits = 4)
```

```
##
## ===============================================================
##                                  Dependent variable:
##                      ---------------------------------------
##                                       birthweight
##                              Model (3)          Model (4)
##                                 (1)                (2)
## ---------------------------------------------------------------
## smoker                     -175.3769***       -176.9589***
##                             (26.8268)          (27.3314)
##
## alcohol                     -21.0835           -14.7583
##                             (72.9921)          (72.9074)
##
## nprevist                    29.6025***         29.7751***
##                             (3.5827)           (3.5973)
##
## unmarried                  -187.1332***       -199.3195***
##                             (27.6772)          (30.9943)
##
## age                                            -2.4935
##                                                (2.4451)
##
```

```
## educ                                                0.2380
##                                                    (5.5328)
##
## Constant                    3,134.4000***          3,199.4260***
##                               (44.1486)              (90.6361)
##
## ------------------------------------------------------------------------
## Observations                   3,000                  3,000
## R2                            0.0886                 0.0890
## Adjusted R2                   0.0874                 0.0872
## Residual Std. Error   565.6975 (df = 2995)   565.7600 (df = 2993)
## F Statistic         72.7930*** (df = 4; 2995) 48.7410*** (df = 6; 2993)
## ========================================================================
## Note:                                        *p<0.1; **p<0.05; ***p<0.01
```

The coefficient on *Smoker* in model (4) is very similar to its value in model (3).