Problem Set 7

Junwoo Yang

May 28, 2021

Exercise 7.4

Using the regression results in column (3):

(a) Are there any important regional differences? Use an appropriate hypothesis test to explain your answer.

(b) Juan is a 32-year-old male high school graduate from the North. Mel is a 32-year-old male high school graduate from the West. Ari is a 32-year-old male high school graduate from the East.

(i) Construct a 95% confidence interval for the difference in expected earnings between Juan and Mel.

(ii) Explain how you would construct a 95% confidence interval for the difference in expected earnings between Juan and Ari.

Answer. (a) We need to test joint hypothesis

$$H_0 \colon \beta_4 = \beta_5 = \beta_6{}^1 = 0 \quad \text{vs.} \quad H_1 \colon \exists j \in \{4,5,6\} \text{ s.t. } \beta_j \neq 0.$$

With assuming that $\hat{\rho}^2_{t_i,t_j} = 0$, we can simply use the F-statistics with

$$F = \frac{1}{3}(t_4^2 + t_5^2 + t_6^2) = \frac{1}{3}\Big(\Big(\frac{\hat{\beta}_4}{SE(\hat{\beta}_4)}\Big)^2 + \Big(\frac{\hat{\beta}_5}{SE(\hat{\beta}_5)}\Big)^2 + \Big(\frac{\hat{\beta}_6}{SE(\hat{\beta}_6)}\Big)^2\Big) = 12.57971,$$

but exact F-statistic for H_0 was already given as 21.87 in table of results, which is asymptotically distributed $\chi^2_3/3$. The 1% critical value from $\chi^2_3/3 = F_{3,\infty}$ is 3.78 which is much lower than 21.87. Thus, the null hypothesis that the regional effects are zero is rejected at 1% significance level. Namely, there are significant differences between regions.

(b) (i) The difference between Juan and Mel is only region which is directly written as β_4. The 95% confidence interval for β_4 is

$$(\hat{\beta}_4 - z_{0.05}\, SE(\hat{\beta}_4),\ \hat{\beta}_4 + z_{0.05}\, SE(\hat{\beta}_4)) = (0.1141404,\ 0.2358596).$$

(ii) The expected difference between Juan and Ari is $\beta_4 - \beta_6$. Once either *North* or *East* is setted for baseline group of region categories, the confidence interval for the difference at any α can be constructed similarly to (i).

Exercise 7.6

In all of the regressions in the previous Exercises, the coefficient of High school is positive, large, and statistically significant. Do you believe this provides strong statistical evidence of the high returns to schooling in the labor market?

Answer. If that is true, this means academic discrimination. However, it is also important to control the characteristics of workers that can affect productivity. If high school graduates are in a society where employment opportunities are guaranteed rather than those who do not, high school graduates will have a relatively higher career year, and low-educated workers will eventually have to accept the wage difference. Therefore, it is premature to conclude academic discrimination based solely on these results of three regressions.

[1] I replaced *East* (X_7) with X_6.

Exercise 7.9

Consider the regression model $Y_i = \beta_0+\beta_1 X_{1i}+\beta_2 X_{2i}+u_i$. Use approach 2 from Section 7.3 to transform the regression so that you can use a t-statistic to test

(a) $\beta_1 = \beta_2$.

(b) $\beta_1 + 2\beta_2 = 0$.

(c) $\beta_1 + \beta_2 = 1$.

Answer. (a) Transform the original regression into

$$Y_i = \beta_0 + (\beta_1 - \beta_2)X_{1i} + \beta_2(X_{1i} + X_{2i}) + u_i = \beta_0 + \gamma X_{1i} + \beta_2 W_i + u_i,$$

and test the hypothesis $H_0\colon \beta_1 - \beta_2 = \gamma = 0$ vs. $H_1\colon \gamma \neq 0$ which is just a single coefficient.

(b) Similarly, transform the regression to

$$Y_i = \beta_0 + (\beta_1 + 2\beta_2)X_{1i} + \beta_2(X_{2i} - 2X_{1i}) + u_i = \beta_0 + \gamma X_{1i} + \beta_2 W_i + u_i,$$

and test whether $\gamma = 0$.

(c) Similarly, put

$$Y_i = \beta_0 + (\beta_1 + \beta_2)X_{1i} + \beta_2(X_{2i} - X_{1i}) + u_i = \beta_0 + \gamma X_{1i} + \beta_2 W_i + u_i,$$

and test whether $\gamma = 1$. Alternatively, redefine the dependent variable of the regression

$$Y_i - X_{1i} = \beta_0 + (\beta_1 + \beta_2 - 1)X_{1i} + \beta_2(X_{2i} - X_{1i}) + u_i = \beta_0 + \gamma X_{1i} + \beta_2 W_i + u_i,$$

and test whether $\gamma = 0$.

Exercise 6.12

A school district undertakes an experiment to estimate the effect of class size on test scores in second-grade classes. The district assigns 50% of its previous year's first graders to small second-grade classes (18 students per classroom) and 50% to regular-size classes (21 students per classroom). Students new to the district are handled differently: 20% are randomly assigned to small classes and 80% to regular-size classes. At the end of the second-grade school year, each student is given a standardized exam. Let Y_i denote the exam score for the i-th student, X_i denote a binary variable that equals 1 if the student is assigned to a small class, and W_i denote a binary variable that equals 1 if the student is newly enrolled. Let β_1 denote the causal effect on test scores of reducing class size from regular to small.

(a) Consider the regression $Y_i = \beta_0+\beta_1 X_i+u_i$. Do you think that $E(u_i|X_i) = 0$? Is the OLS estimator of β_1 unbiased and consistent? Explain.

(b) Consider the regression $Y_i = \beta_0 + \beta_1 X_i + \beta_2 W_i + u_i$. Do you think that $E(u_i|X_i, W_i)$ depends on X_i? Is the OLS estimator of β_1 unbiased and consistent? Explain. Do you think that $E(u_i|X_i, W_i)$ depends on W_i? Will the OLS estimator of β_2 provide an unbiased and consistent estimate of the causal effect of transferring to a new school (that is, being a newly enrolled student)? Explain.

Answer. (a) Treatment (assignment to small classes) was not randomly assigned in the population (the continuing and newly-enrolled students) because of the difference in the proportion of the treated among the continuing and newly-enrolled students. Thus, the treatment indicator X is correlated with W. If newly-enrolled students perform systematically differently on standardized tests than continuing students (perhaps because of adjustment to a new school), then this becomes part of the error term u. This leads to correlation between X and u, so that $E(u|X) \neq 0$. Because $E(u|X) \neq 0$, the $\hat{\beta}_1$ is biased and inconsistent.

(b) Because treatment was randomly assigned conditional on enrollment status (continuing or newly-enrolled), $E(u|X, W)$ will not depend on X. This means that the assumption of conditional mean independence is satisfied, i.e., $E(u|X, W) = E(u|W)$, and $\hat{\beta}_1$ is unbiased and consistent. However, because W was not randomly assigned (newly-enrolled students may, on average, have attributes other than being newly enrolled that affect test scores), $E(u|X, W)$ may depend on W, so that $\hat{\beta}_2$ may be biased and inconsistent.

Empirical Exercise 7.1

Answer the following questions.

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

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

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

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

(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?

(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?

Answer. (a) -253.228, -217.580, 175.377.

(b) $(-305.776, -200.681)$, $(-268.750, -166.410)$, $(-227.956, -122.797)$.

(c) Since there is another significant regressor `nprevist` in regression (2), the coefficient on `smoker` in regression (1) suffer from omitted variable bias.

(d) With the same logic, because of another significant regressor `unmarried` in regression (3), the coefficient on `smoker` in regression (2) would suffer from omitted variable bias.

(e) (i) $(-241.379, -132.887)$.

  (ii) Yes, it is. The coefficient is included in the 95% confidence interval, and equivalently the p-value is very low.

  (iii) With other variables fixed, infants of single mothers 187.133g lighter, in average, than infants of married mothers. This is the largest relative to the coefficients of other variables, indicating that marital status has a relatively large effect on the infant's weight.

  (iv) Although the result shows relatively large differences in infant weight depending on marital status, it is difficult to conclude that it is a real causal effect. Suppose nutrition is a determinant of an infant's weight. If married mothers in data belonged to more nutritious environment during pregnancy, marital status would not be a determinant of infant weight.

(f) Removing insignificant `alcohol`, and further considering `educ` and `age`, regression (4) consists only of significant regressors. Thus, $(-228.5775, -123.8470)$ is a reasonable 95% confidence interval.

Appendix: R code

```r
library(readxl)
bs <- read_excel("birthweight_smoking/birthweight_smoking.xlsx")

# (a) ~ (e)
ols1 <- lm(birthweight ~ smoker, data=bs)
ols2 <- lm(birthweight ~ smoker + alcohol + nprevist, data=bs)
ols3 <- lm(birthweight ~ smoker + alcohol + nprevist + unmarried, data=bs)

library(sandwich)
Cov1 <- vcovHC(ols1, type="HC1")
Cov2 <- vcovHC(ols2, type="HC1")
Cov3 <- vcovHC(ols3, type="HC1")

library(dplyr)
selist <- list(Cov1 %>% diag() %>% sqrt(), Cov2 %>% diag() %>% sqrt(),
Cov3 %>% diag() %>% sqrt())

library(stargazer)
stargazer(ols1, ols2, ols3, type="text", se=selist, intercept.bottom=FALSE, omit.stat=c("f"),
title="Results of three regressions with heteroskedasticity-robust standard errors")

##
## Results of three regressions with heteroskedasticity-robust standard errors
## ================================================================================
##                                        Dependent variable:
##                     ---------------------------------------------------------
##                                             birthweight
##                              (1)                 (2)                 (3)
## --------------------------------------------------------------------------------
## Constant                3,432.060***        3,051.249***        3,134.400***
##                           (11.891)            (43.714)            (44.149)
##
## smoker                   -253.228***         -217.580***         -175.377***
##                           (26.810)            (26.108)            (26.827)
##
## alcohol                                        -30.491             -21.083
##                                               (72.597)            (72.992)
##
## nprevist                                      34.070***           29.603***
##                                                (3.608)             (3.583)
##
## unmarried                                                        -187.133***
##                                                                   (27.677)
##
## --------------------------------------------------------------------------------
## Observations               3,000               3,000               3,000
## R2                         0.029               0.073               0.089
## Adjusted R2                0.028               0.072               0.087
## Residual Std. Error 583.730 (df = 2998) 570.471 (df = 2996) 565.698 (df = 2995)
## ================================================================================
## Note:                                            *p<0.1; **p<0.05; ***p<0.01
```

```
stargazer(ols1, ols2, ols3, type="text", se=selist, intercept.bottom=FALSE, omit.stat=c("f"),
ci=TRUE, ci.level=0.95, report="vc*stp",
title="with 95% confidence intervals, t-statistics, p-values")
```

```
##
## with 95% confidence intervals, t-statistics, p-values
## ================================================================================================
##                                              Dependent variable:
##                      ------------------------------------------------------------------------
##                                                  birthweight
##                                (1)                     (2)                      (3)
## ------------------------------------------------------------------------------------------------
## Constant                   3,432.060***            3,051.249***             3,134.400***
##                       (3,408.755, 3,455.365) (2,965.570, 3,136.927) (3,047.870, 3,220.930)
##                            t = 288.638             t = 69.800               t = 70.997
##                             p = 0.000               p = 0.000                p = 0.000
##
## smoker                     -253.228***             -217.580***              -175.377***
##                         (-305.776, -200.681)    (-268.750, -166.410)     (-227.956, -122.797)
##                            t = -9.445              t = -8.334               t = -6.537
##                             p = 0.000               p = 0.000                p = 0.000
##
## alcohol                                             -30.491                  -21.083
##                                                  (-172.778, 111.796)      (-164.145, 121.978)
##                                                     t = -0.420               t = -0.289
##                                                     p = 0.675                p = 0.773
##
## nprevist                                            34.070***                29.603***
##                                                   (26.998, 41.142)         (22.581, 36.625)
##                                                     t = 9.442                t = 8.263
##                                                     p = 0.000                p = 0.000
##
## unmarried                                                                    -187.133***
##                                                                           (-241.379, -132.887)
##                                                                              t = -6.761
##                                                                              p = 0.000
##
## ------------------------------------------------------------------------------------------------
## Observations                 3,000                   3,000                    3,000
## R2                           0.029                   0.073                    0.089
## Adjusted R2                  0.028                   0.072                    0.087
## Residual Std. Error  583.730 (df = 2998)    570.471 (df = 2996)      565.698 (df = 2995)
## ================================================================================================
## Note:                                                          *p<0.1; **p<0.05; ***p<0.01
```

```r
# (f)
ols4 <- lm(birthweight ~ smoker + nprevist + unmarried, data=bs)
ols5 <- lm(birthweight ~ smoker + nprevist + unmarried + educ, data=bs)
ols6 <- lm(birthweight ~ smoker + nprevist + unmarried + age, data=bs)
Cov4 <- vcovHC(ols4, type="HC1")
Cov5 <- vcovHC(ols5, type="HC1")
Cov6 <- vcovHC(ols6, type="HC1")
selist2 <- append(selist, list(Cov4 %>% diag() %>% sqrt(), Cov5 %>% diag() %>% sqrt(),
Cov6 %>% diag() %>% sqrt()))
stargazer(ols1, ols2, ols3, ols4, ols5, ols6, type="text", se=selist2, intercept.bottom=FALSE,
omit.stat=c("n","ser","f"), title="Table for (f)")
```

```
##
## Table for (f)
## ================================================================================================
##                                          Dependent variable:
##             ---------------------------------------------------------------------------
##                                              birthweight
##                 (1)          (2)          (3)          (4)          (5)          (6)
## ------------------------------------------------------------------------------------------------
## Constant    3,432.060*** 3,051.249*** 3,134.400*** 3,133.957*** 3,158.304*** 3,202.003***
##               (11.891)     (43.714)     (44.149)     (44.113)     (79.999)     (76.405)
##
## smoker      -253.228***  -217.580***  -175.377***  -176.212***  -177.841***  -177.741***
##               (26.810)     (26.108)     (26.827)     (26.707)     (27.241)     (26.855)
##
## alcohol                    -30.491      -21.083
##                            (72.597)     (72.992)
##
## nprevist                  34.070***    29.603***    29.623***    29.771***    29.807***
##                            (3.608)      (3.583)      (3.582)      (3.600)      (3.585)
##
## unmarried                             -187.133***  -187.259***  -190.016***  -199.727***
##                                         (27.677)     (27.669)     (29.027)     (30.614)
##
## educ                                                                -1.940
##                                                                    (5.229)
##
## age                                                                            -2.490
##                                                                               (2.299)
##
## ------------------------------------------------------------------------------------------------
## R2             0.029        0.073        0.089        0.089        0.089        0.089
## Adjusted R2    0.028        0.072        0.087        0.088        0.087        0.088
## ================================================================================================
## Note:                                                        *p<0.1; **p<0.05; ***p<0.01
```

```r
library(lmtest)
coeftest(ols4, vcov.=Cov4)
```

```
##
## t test of coefficients:
##
##              Estimate Std. Error t value  Pr(>|t|)
## (Intercept) 3133.9572    44.1128 71.0442 < 2.2e-16 ***
## smoker      -176.2122    26.7067 -6.5981 4.907e-11 ***
## nprevist      29.6231     3.5821  8.2697 < 2.2e-16 ***
## unmarried   -187.2587    27.6687 -6.7679 1.566e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
coefci(ols4, vcov.=Cov4)["smoker",]
```

```
##     2.5 %    97.5 %
## -228.5775 -123.8470
```