Chapter 5 Problem Set

Tianwei Liu

10/19/2019

3. Do cell phones distract drivers and cause accidents? Worried that this is happening, many states recently have passed legislation to reduce distracted driving. Fourteen states now have laws making handheld cell phone use while driving illegal, and 44 states have banned texting while driving. This problem looks more closely at the relationship between cell phones and traffic fatalities. Table 5.11 describes the variables in the data set Cellphone_2012_homework.dta.

a. While we don't know how many people are using their phones while driving, we can find the number of cell phone subscriptions in a state (in thousands). Estimate a bivariate model with traffic deaths as the dependent variable and number of cell phone subscriptions as the independent variable. Briefly discuss the results. Do you suspect endogeneity? If so, why?

```
load("Ch5_Exercise3_Cell_phone_subscriptions.RData")

reg1 <- lm(numberofdeaths ~ cell_subscription, data = dta)
summary(reg1)
```

```
##
## Call:
## lm(formula = numberofdeaths ~ cell_subscription, data = dta)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -844.04 -123.11  -56.48  151.64 1036.14
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       1.240e+02  5.464e+01   2.269   0.0278 *
## cell_subscription 9.115e-02  6.095e-03  14.955   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 287.1 on 48 degrees of freedom
## Multiple R-squared:  0.8233, Adjusted R-squared:  0.8196
## F-statistic: 223.6 on 1 and 48 DF,  p-value: < 2.2e-16
```

The estimated coefficient on number of cell phone subscription is 0.09, meaning that one more cellphone subscription in a state tends to increase number of deaths by 0.09. This coefficient is statistically significant because the t-value is much greater than 2.

I do suspect endogeneity. Both number of deaths and number of cell phone subscription is associated with the total population in a state. More populous the state is, the more deaths and more cellphone subscription. Therefore, there is endogeneity in this model.

b. Add population to the model. What happens to the coefficient on cell phone subscriptions? Why?

```
reg2 <- lm(numberofdeaths ~ cell_subscription + population, data = dta)
summary(reg2)
```

```
##
## Call:
## lm(formula = numberofdeaths ~ cell_subscription + population,
##     data = dta)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -811.0 -128.7  -47.8  138.5  882.2
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        1.139e+02  4.991e+01   2.283  0.02702 *
## cell_subscription -2.109e-01  9.230e-02  -2.285  0.02689 *
## population         2.909e-04  8.873e-05   3.278  0.00197 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 261.7 on 47 degrees of freedom
## Multiple R-squared:  0.8562, Adjusted R-squared:  0.8501
## F-statistic: 139.9 on 2 and 47 DF,  p-value: < 2.2e-16
```

The estimated coefficient on cell phone subscription becomes negative and less statistically significant after accounting for population. In (a), since population influences the number of deaths and is correlated with cell_subscription, it is an omitted variable. When we have an omitted variable, our estimate of beta1_hat is biased. Therefore, when we account for the omitted variable, the bias is reduced or removed. This is why there's a change in the estimated coefficient on cell_subscription.

c. Add total miles driven to the model. What happens to the coefficient on cell phone subscriptions? Why?

```
reg3 <- lm(numberofdeaths ~ cell_subscription + population + total_miles_driven, data =
dta)
summary(reg3)
```

```
##
## Call:
## lm(formula = numberofdeaths ~ cell_subscription + population +
##     total_miles_driven, data = dta)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -555.98  -92.75  -12.18   60.67  788.37
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         4.346e+00  4.108e+01   0.106    0.916
## cell_subscription   2.465e-03  7.671e-02   0.032    0.975
## population         -7.422e-05  8.821e-05  -0.841    0.404
## total_miles_driven  1.883e-02  3.018e-03   6.240 1.27e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 194.7 on 46 degrees of freedom
## Multiple R-squared:  0.9221, Adjusted R-squared:  0.917
## F-statistic: 181.5 on 3 and 46 DF,  p-value: < 2.2e-16
```

The coefficient on cell phone subscription becomes not significant. In this case, total miles driven seems like an irrelevant variable. How many miles one state's citizens drives does not appear to explain the number of deaths. When we include irrelevant variables, the variance of the estimated coefficient on cell_subscription becomes higher and therefore the t-value becomes lower; and it is not statistically significant.

d. Based on the model in part (c), calculate the variance inflation factor for population and total miles driven. Why are they different? Discuss implications of this level of multicollinearity for the coefficient estimates and the precision of the coefficient estimates.

```
vif(reg3)
```

```
##  cell_subscription         population total_miles_driven
##           344.3690           492.7790            43.0868
```

The VIF for population is 492.779. VIF for total miles driven is 43.0868, indicating a high level of multicollinearity. They are different because
population is highly correlated with number of cellphone subscription, as cellphone is almost universal in the entire population in the US. A high collinearity also causes the variance of the estimated coefficients to be higher. This is also why coefficients become insignificant.

4. What determines how much drivers are fined if they are stopped for speeding? Do demographics like age, gender, and race matter? To answer this question, we'll investigate traffic stops and citations in Massachusetts using data from Makowsky and Stratmann (2009). Even though state law sets a formula for tickets based on how fast a person was driving, police officers in practice often deviate from the formula. Table 5.12 describes data in speeding_tickets_text.dta that includes information on all traffic stops. An amount for the fine is given only for observations in which the police officer decided to assess a fine.

a. Estimate a bivariate OLS model in which ticket amount is a function of age. Is age statistically significant? Is endogeneity possible?

```
load("Ch5_Exercise4_Speeding_tickets.RData")
reg4 <- lm(Amount ~ Age, data = dta)
summary(reg4)
```

```
##
## Call:
## lm(formula = Amount ~ Age, data = dta)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -123.21  -46.58   -5.92   32.55  600.24
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 131.70665    0.88649  148.57   <2e-16 ***
## Age          -0.28927    0.02478  -11.68   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 56.13 on 31672 degrees of freedom
##   (36683 observations deleted due to missingness)
## Multiple R-squared:  0.004286,   Adjusted R-squared:  0.004254
## F-statistic: 136.3 on 1 and 31672 DF,  p-value: < 2.2e-16
```

The t-stat for Age is -11.68; coefficient on age is statistically significant. Endogeneity is possible. Let’s consider MPH over the speed limit. The larger the MPH over speed limit, the more the fine will be. Old people may drive more slowly than younger drivers, so their MPH over is likely smaller.

b. Estimate the model from part (a), also controlling for miles per hour over the speed limit. Explain what happens to the coefficient on age and why.

```
reg5 <- lm(Amount ~ Age + MPHover, data = dta)
summary(reg5)
```

```
##
## Call:
## lm(formula = Amount ~ Age + MPHover, data = dta)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -308.763  -19.783    7.682   25.757  226.295
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.49386    0.95428   3.661 0.000251 ***
## Age          0.02496    0.01760   1.418 0.156059
## MPHover      6.89175    0.03869 178.130  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 39.67 on 31671 degrees of freedom
##   (36683 observations deleted due to missingness)
## Multiple R-squared:  0.5026, Adjusted R-squared:  0.5026
## F-statistic: 1.6e+04 on 2 and 31671 DF,  p-value: < 2.2e-16
```

The coefficient on age becomes positive and not significant. It is because in the model before, MPHover is an omitted variable, and the esimate for coefficient on age is biased.

c. Suppose we had only the first thousand observations in the data set. Estimate the model from part (b), and report on what happens to the standard errors and t statistics when we have fewer observations.

```
dta2 = dta[1:1000,]
reg6 <- lm(Amount ~ Age + MPHover, data = dta2)
summary(reg6)
```

```
##
## Call:
## lm(formula = Amount ~ Age + MPHover, data = dta2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -173.599   -4.623    4.405   24.898  102.575
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -1.7795     8.9592  -0.199    0.843
## Age           0.1813     0.1918   0.945    0.345
## MPHover       6.8566     0.3423  20.033   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 42.89 on 336 degrees of freedom
##   (661 observations deleted due to missingness)
## Multiple R-squared:  0.545,  Adjusted R-squared:  0.5423
## F-statistic: 201.2 on 2 and 336 DF,  p-value: < 2.2e-16
```

Compared with the multivariate model on the entire dataset, the model on the first 1000 observations yields a higher standard error and as a result, a lower t value. This can be understood by the var(beta_j) formula, a larger N will lower the variance.