Ch 5 Exercise 3

Allen Church

10/21/2019

Load data

```
require(knitr)
require(haven)
require(AER)

opts_chunk$set(echo = TRUE)
options(digits = 3)

#Add working directory, assign data to lab variable
data <- load("/Users/allenchurch/Ch5_Exercise3_Cell_phone_subscriptions.RData")
```

3a. Create bivariate model with traffic deaths as dependent variable and number of cell phone subscriptions as independent variable. Do you suspect endogeneity, if so why?

```
ols1 <- lm(dta$numberofdeaths ~ dta$cell_subscription)
summary(ols1)
```

```
##
## Call:
## lm(formula = dta$numberofdeaths ~ dta$cell_subscription)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -844.0 -123.1  -56.5  151.6 1036.1
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)             1.24e+02   5.46e+01    2.27    0.028 *
## dta$cell_subscription 9.11e-02   6.09e-03   14.95   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 287 on 48 degrees of freedom
## Multiple R-squared:  0.823,  Adjusted R-squared:  0.82
## F-statistic:  224 on 1 and 48 DF,  p-value: <2e-16
```

The results above indicate endogeneity, since the testimated coefficient is 9.115e-02. As this result seems incredibly high, we suspect endogeneity since changes in this independent variable (cell_subscription) are related to factors in the error term.

3b. Add population to model. What happens to the coefficients on cell phone subscriptions, why?

```
ols2 <- lm(dta$numberofdeaths ~ dta$cell_subscription + dta$population)
summary(ols2)
```

```
##
## Call:
## lm(formula = dta$numberofdeaths ~ dta$cell_subscription + dta$population)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -811.0 -128.7  -47.8  138.5  882.2
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             1.14e+02   4.99e+01    2.28    0.027 *
## dta$cell_subscription -2.11e-01   9.23e-02   -2.28    0.027 *
## dta$population         2.91e-04   8.87e-05    3.28    0.002 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 262 on 47 degrees of freedom
## Multiple R-squared:  0.856,  Adjusted R-squared:  0.85
## F-statistic:  140 on 2 and 47 DF,  p-value: <2e-16
```

After adding population to the model, the coefficient on cell_subscription changes from positive to negative. This indicates that the effects of population coefficient were included in cell_subscription, which indicates that our previous results were endogeneous.

3c. Add total miles driven to the model. What happens to coefficient on cell phone subscriptions, why?

```
ols3 <- lm(dta$numberofdeaths ~ dta$cell_subscription + dta$population + dta$total_miles_driven)
summary(ols3)
```

```
##
## Call:
## lm(formula = dta$numberofdeaths ~ dta$cell_subscription + dta$population +
##     dta$total_miles_driven)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -556.0  -92.7  -12.2   60.7  788.4
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)              4.35e+00   4.11e+01    0.11     0.92
## dta$cell_subscription    2.46e-03   7.67e-02    0.03     0.97
## dta$population          -7.42e-05   8.82e-05   -0.84     0.40
## dta$total_miles_driven   1.88e-02   3.02e-03    6.24  1.3e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 195 on 46 degrees of freedom
## Multiple R-squared:  0.922,  Adjusted R-squared:  0.917
## F-statistic:  182 on 3 and 46 DF,  p-value: <2e-16
```

Adding total miles driven to the model causes the coefficient on cell phone subscriptions to turn from negative to positive, from -2.109e-01 to 2.465e-03. This indicates omitted variable bias, in which the variable

affects the dependent variable (number of deaths) and is correlated with the independent variable (cell phone subscriptions).

3d. Based on 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 coefficient estimates.

```
vif(ols3)
```

```
##  dta$cell_subscription         dta$population dta$total_miles_driven
##                  344.4                  492.8                   43.1
```

The variance inflation factor above represents how much variance is inflated due to multicollinearity. For cell_subscription, the VIF is 344.3690, which suggests cell_subscription is highly related to other independent variables. The VIF for population is 492.7790, which suggests that this variable is the highest related to the other independent variables. The VIF for miles_driven is 43.0868, which while not as high as the other two, is still high enough for us to note uncertainty in our results. High amounts of multicollinearity prevents the producers of an analysis from asserting much about the results, again due to high variance. However, it is important to note that despite the model having multicollinearity, this does not cause bias.