Homework 1 (Part 2) for Marketing Aanlytics

Regression and Endogeneity

Jaskaran Singh Kohli

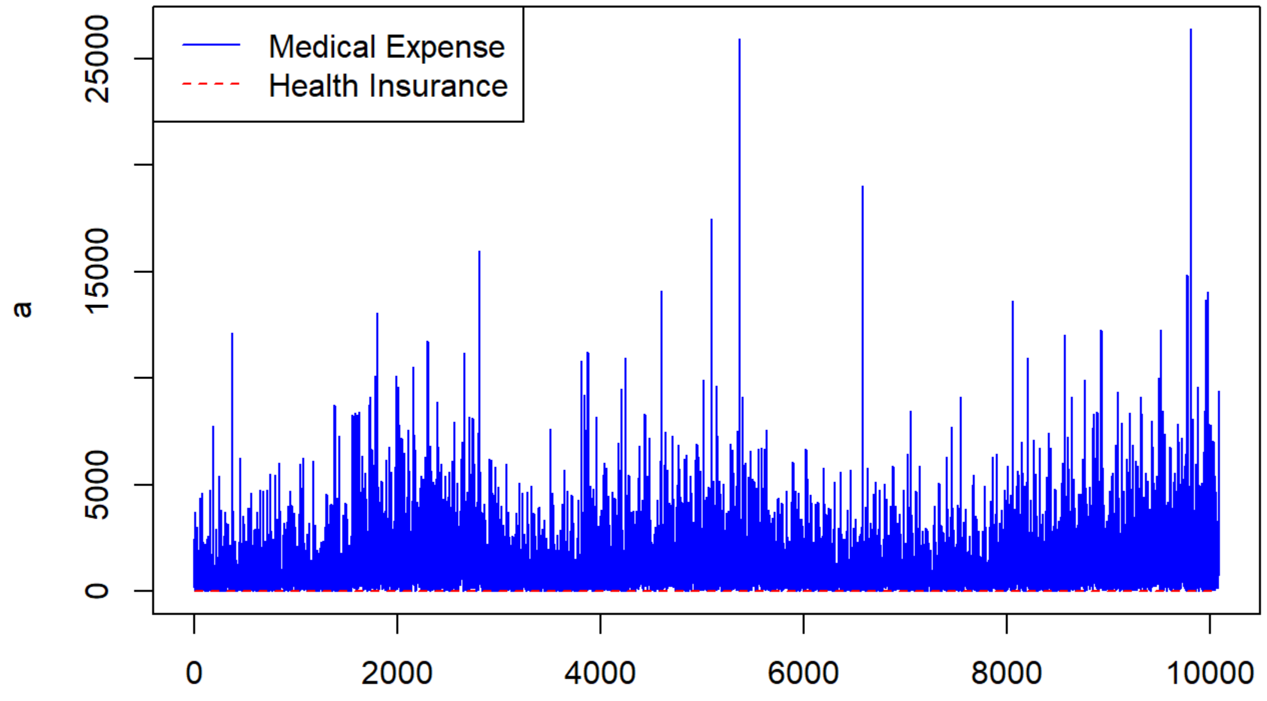
Due on Monday, January 22, 2020

1. Load the data file `health_inclass.csv` , conduct simple regression without correcting for endogeneity, and try to answer the question whether having health insurance leads to higher or lower medical expenses. In this exercise, add more variables from the data, you can create dummy variables, add meaningful interaction variables. Try at least three models (different specifications from the example in class), and find the best one among the three, interpret the model results.

Present all the three model results, and answer the following questions:

1. Based on what metrics did you choose the “best” model?
2. Do you think the endogeneity of the $HealthIns$ variable still exists? Why or why not?

```
df= read.csv("health_inclass.csv", header = TRUE)
a = cbind(df$medexpense,df$healthinsu)
matplot(a,type="l",col=c("blue","red"))
legend('topleft',c("Medical Expense","Health Insurance"),lty=1:2,col=c("blue","red"),cex=1)
```



```
attach(df)
mean(df$age)#Average age of people in the dataset
```

```
## [1] 75.05174
```

```
Y = log(df$medexpense)
logincome = log(df$income)
X = cbind(healthinsu,illnesses,age*vegood*blackhisp,female)#Key Independent and Control Variable
s
med.olsreg <- lm(Y~X)
summary(med.olsreg)
```

```
##
## Call:
## lm(formula = Y ~ X)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.2794 -0.6775  0.1463  0.8527  3.7326
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.6060715  0.0278964 200.961  < 2e-16 ***
## Xhealthinsu  0.0915852  0.0254835   3.594 0.000327 ***
## Xillnesses   0.4371093  0.0095466  45.787  < 2e-16 ***
## X           -0.0019178  0.0009057  -2.117 0.034252 *
## Xfemale      0.0551213  0.0251024   2.196 0.028125 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.237 on 10084 degrees of freedom
## Multiple R-squared:  0.1753, Adjusted R-squared:  0.175
## F-statistic:   536 on 4 and 10084 DF,  p-value: < 2.2e-16
```

Model 1

Observation

Adjusted R Square = *0.175* P Value for every parameter is less than *0.05* . So this indicate that the estimate is statistically insignificant. Standard Error of all the estimates are very less , ndicating the data is certain . t Stat - Value of all the perimeters is greater than *2* . Interaction Variable - I am checking the combine effect of age of a black hispanic with very good health condition on the the medical expenses . So the model shows that their combine effect decreases the medical expenses .

```
Z = cbind(healthinsu,illnesses,age*poor*female)#Key Independent and Control Variables
med.olsreg2 <- lm(Y~Z) #Model 2
summary(med.olsreg2)
```

```
## 
## Call:
## lm(formula = Y ~ Z)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -6.2409 -0.6736  0.1496  0.8497  3.7632 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 5.630663   0.023641 238.175  < 2e-16 ***
## Zhealthinsu 0.090182   0.025372   3.554 0.000381 ***
## Zillnesses  0.434456   0.009621  45.158  < 2e-16 ***
## Z           0.002929   0.000886   3.306 0.000950 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 1.237 on 10085 degrees of freedom
## Multiple R-squared:  0.1755, Adjusted R-squared:  0.1752 
## F-statistic: 715.4 on 3 and 10085 DF,  p-value: < 2.2e-16
```

Model 2

Observation

Adjusted R Square = *0.1752* P Value for every parameter is less than *0.05* . So this indicate that the estimate is statistically insignificant. Standard Error of all the estimates are very less , ndicating the data is certain . t Stat - Value of all the perimeters is greater than *2* . Interaction Variable - I am checking the combine effect of age of a person with poor health condition on the the medical expenses.So the model shows that their combine effect increases the medical expenses .

```
W = cbind(healthinsu,illnesses*priolist,age,priolist,illnesses) #Key Independent and Control Var
iables
med.olsreg3 <- lm(Y~W) #Model 3
summary(med.olsreg3)
```

```
##
## Call:
## lm(formula = Y ~ W)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.3244 -0.6649  0.1440  0.8322  4.3670
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.340216   0.143980  37.090  < 2e-16 ***
## Whealthinsu  0.082028   0.025317   3.240   0.0012 **
## W           -0.289322   0.040354  -7.170 8.05e-13 ***
## Wage        -0.003869   0.001847  -2.094   0.0362 *
## Wpriolist    0.805165   0.048767  16.510  < 2e-16 ***
## Willnesses   0.653240   0.038991  16.754  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.22 on 10083 degrees of freedom
## Multiple R-squared:  0.1976, Adjusted R-squared:  0.1972
## F-statistic: 496.6 on 5 and 10083 DF,  p-value: < 2.2e-16
```

Model 3

Observation

Adjusted R Square = *0.1972* P Value for every parameter is less than *0.05* . So this indicate that the estimate is statistically insignificant. Standard Error of all the estimates are very less , ndicating the data is certain . t Stat - Value of all the perimeters is greater than *2* . Interaction Variable - I am checking the combine effect of illness falling on priority list on the the medical expenses . So the model shows that their combine effect decreases the medical expenses .

Summary

I created 3 models and with each and every model there is an ambiguity in the dependent and Key independent variable . Because having a health insurance should actually decrease the medical expenses which is not happening . This means that their is correlation between health insurance and medical expenses with the error term . For the Second Question I am taking model 3 as it is giving me the best results with maximum adjusted R square . We'll be taking SSI ratio (ratio of supplemental social income, over total household income. The highest value can be one, meaning the person has no income other than the help from the federal government) as our instrument variable.

Part II Endogeneity and 2SLS

2. Suppose the $HealthIns$ is still endogenous, even with your "best" model, use `SSIRatio` variable as your instrument, and conduct the following exercises

1. Use `ivreg()` conduct the 2SLS estimates for your “best” model, while correcting for endogeneity of the $HealthIns$ variable.
2. Compare the results from this model with those from the simple OLS approach, interms of model fit, parameter interpretations, and your answers to the question “whether having health insurance leads to higher or lower medical expnses.”

```
Y1 <- log(medexpense)
Y2 <- healthinsu
X1 <- cbind(illnesses,age*poor,blackhisp)
X2 <- cbind(ssiratio)

#install.packages("AER")
library(AER)
```

```
## Warning: package 'AER' was built under R version 3.6.2
```

```
## Loading required package: car
```

```
## Warning: package 'car' was built under R version 3.6.2
```

```
## Loading required package: carData
```

```
## Loading required package: lmtest
```

```
## Warning: package 'lmtest' was built under R version 3.6.2
```

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

```
## Warning: package 'zoo' was built under R version 3.6.2
```

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

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

```
## Loading required package: sandwich
```

```
## Warning: package 'sandwich' was built under R version 3.6.2
```

```
## Loading required package: survival
```

```
# 2SLS estimation
med.ivreg <- ivreg(Y1 ~ Y2 + X1 | X1 + X2)
summary(med.ivreg,, diagnostics=TRUE)
```

```
##
## Call:
## ivreg(formula = Y1 ~ Y2 + X1 | X1 + X2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.4458 -0.6921  0.1499  0.8532  3.6326
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.8382333  0.0538419 108.433  < 2e-16 ***
## Y2          -0.3823049  0.1242855  -3.076   0.0021 **
## X1illnesses  0.4335876  0.0098308  44.105  < 2e-16 ***
## X1           0.0026910  0.0006827   3.942 8.14e-05 ***
## X1blackhisp -0.1897088  0.0353884  -5.361 8.47e-08 ***
##
## Diagnostic tests:
##                    df1   df2 statistic  p-value
## Weak instruments     1 10084    453.94  < 2e-16 ***
## Wu-Hausman           1 10083     15.18 9.82e-05 ***
## Sargan               0    NA        NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.256 on 10084 degrees of freedom
## Multiple R-Squared: 0.1506,  Adjusted R-squared: 0.1503
## Wald test: 528.3 on 4 and 10084 DF,  p-value: < 2.2e-16
```

```
ests=cbind(med.olsreg3$coefficients,med.ivreg$coefficients)
```

```
## Warning in cbind(med.olsreg3$coefficients, med.ivreg$coefficients): number
## of rows of result is not a multiple of vector length (arg 2)
```

```
colnames(ests) = c('OLS', '2SLS-ivreg')
ests
```

```
##                      OLS   2SLS-ivreg
## (Intercept)  5.340215617  5.838233262
## Whealthinsu  0.082028105 -0.382304918
## W           -0.289322189  0.433587592
## Wage        -0.003868906  0.002690994
## Wpriolist    0.805164986 -0.189708848
## Willnesses   0.653239822  5.838233262
```

```
Medical_Insurance_Parameter = ests[2,]
Medical_Insurance_Parameter*100
```

```
##        OLS 2SLS-ivreg
##    8.20281  -38.23049
```

2SLS show a 45.8 % decrease in medical expenses when we have a medical insurance .

```
stderrs <- cbind(summary(med.olsreg3)$coefficients[,2],
            summary(med.ivreg)$coefficients[,2])
```

```
## Warning in cbind(summary(med.olsreg3)$coefficients[, 2], summary(med.ivreg)
## $coefficients[, : number of rows of result is not a multiple of vector
## length (arg 2)
```

```
colnames(stderrs)=c('OLS', '2SLS-ivreg')
stderrs
```

```
##                     OLS   2SLS-ivreg
## (Intercept) 0.143979647 0.0538418641
## Whealthinsu 0.025317145 0.1242855114
## W           0.040353731 0.0098308024
## Wage        0.001847242 0.0006826553
## Wpriolist   0.048767346 0.0353883673
## Willnesses  0.038991045 0.0538418641
```

```
#Comparing the first and second columns, we can notice that although the standard errors from th
e other
#parameters are quite close, the standard error for the endogenous variable are very different
```

2SLS shows a minor increase in standard error

Final Interpretion

It seems that Health Insurance was an endogenous parameter and we removed the endogenity with by running 2SLS model . All the parameters in the new model are significant with a minor increase in standard error and finally estimate we are getting is " Your medical expense will decrease by 45.8% when we " .Wu-Hausman test values are also significant . So in Conclusion , I *will recommend* people to take medical insurance