Homework 1 Part 2 for Marketing Aanlytics

Regression and Endogeneity

Harsh Tandon

Due on Wednesday, January 22, 2020

Part II Endogeneity and 2SLS

Question 1

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?

Set working directory, Import libraries, Read data, Print summary statistics

```
setwd("D:/2nd Qtr Study Material/Marketing Analytics/Lecture 2 Regression/Lecture 2 Regressio
n")
#Install required packages
#install.packages("stargazer")
#install.packages("lmtest")
#install.packages("VIF")
#install.packages("AER")
library(stargazer)
```

```
##
## Please cite as:
```

```
##  Hlavac, Marek (2018). stargazer: Well-Formatted Regression and Summary Statistics Tables.
```

```
##  R package version 5.2.2. https://CRAN.R-project.org/package=stargazer
```

```
library(ggplot2)
library(graphics)
library(lmtest)
```

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

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

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

```
library(VIF)
library(AER)
```

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

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

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:VIF':
##
##     vif
```

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

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

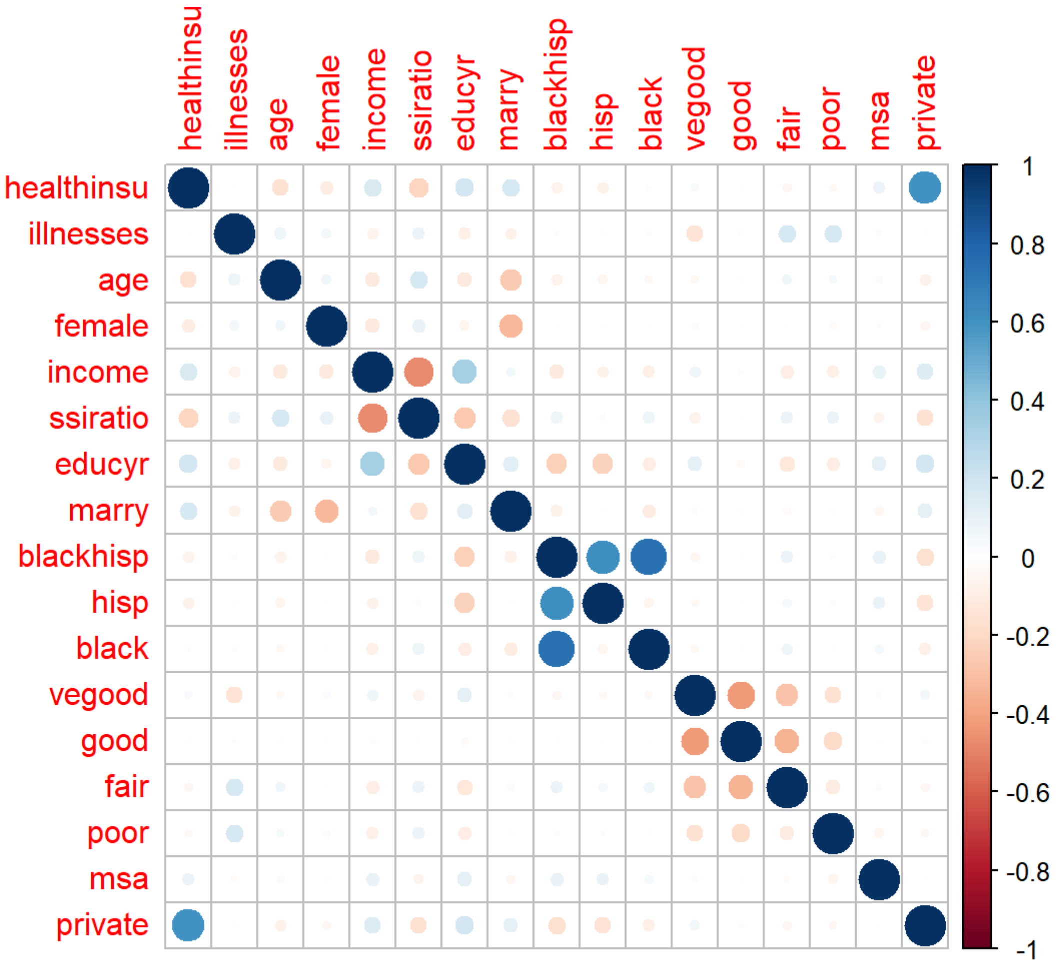
```
data = read.csv("health_inClass.csv", header = TRUE) #read data

stargazer(data, type="text", median=TRUE, iqr=TRUE, digits=2, title="Descriptive Statistics")
#print summary statistics
```

```
##
## Descriptive Statistics
## =========================================================================
## Statistic    N      Mean   St. Dev.  Min  Pctl(25) Median Pctl(75)  Max
## -------------------------------------------------------------------------
## indid      10,089 5,045.00 2,912.59   1    2,523    5,045    7,567   10,089
## medexpense 10,089 1,287.56 1,530.42   1     311      795     1,685   26,375
## healthinsu 10,089   0.38     0.49     0      0       0       1       1
## illnesses  10,089   1.86     1.29     0      1       2       3       9
## age        10,089  75.05     6.68    65      70      74      80      91
## female     10,089   0.58     0.49     0      0       1       1       1
## income     10,089  22.12    21.74   0.001   9.32    15.54   27.52   312.46
## ssiratio   10,089   0.54     0.37    0.00    0.24    0.50    0.91    9.25
## educyr     10,089  11.79     3.24     0      10      12      14      17
## marry      10,089   0.56     0.50     0      0       1       1       1
## blackhisp  10,089   0.16     0.37     0      0       0       0       1
## hisp       10,089   0.07     0.25     0      0       0       0       1
## black      10,089   0.10     0.30     0      0       0       0       1
## vegood     10,089   0.26     0.44     0      0       0       1       1
## good       10,089   0.34     0.47     0      0       0       1       1
## fair       10,089   0.19     0.39     0      0       0       0       1
## poor       10,089   0.07     0.25     0      0       0       0       1
## msa        10,089   0.73     0.44     0      0       1       1       1
## private    10,089   0.63     0.48     0      0       1       1       1
## priolist   10,089   0.87     0.34     0      1       1       1       1
## -------------------------------------------------------------------------
```

Check for correlation

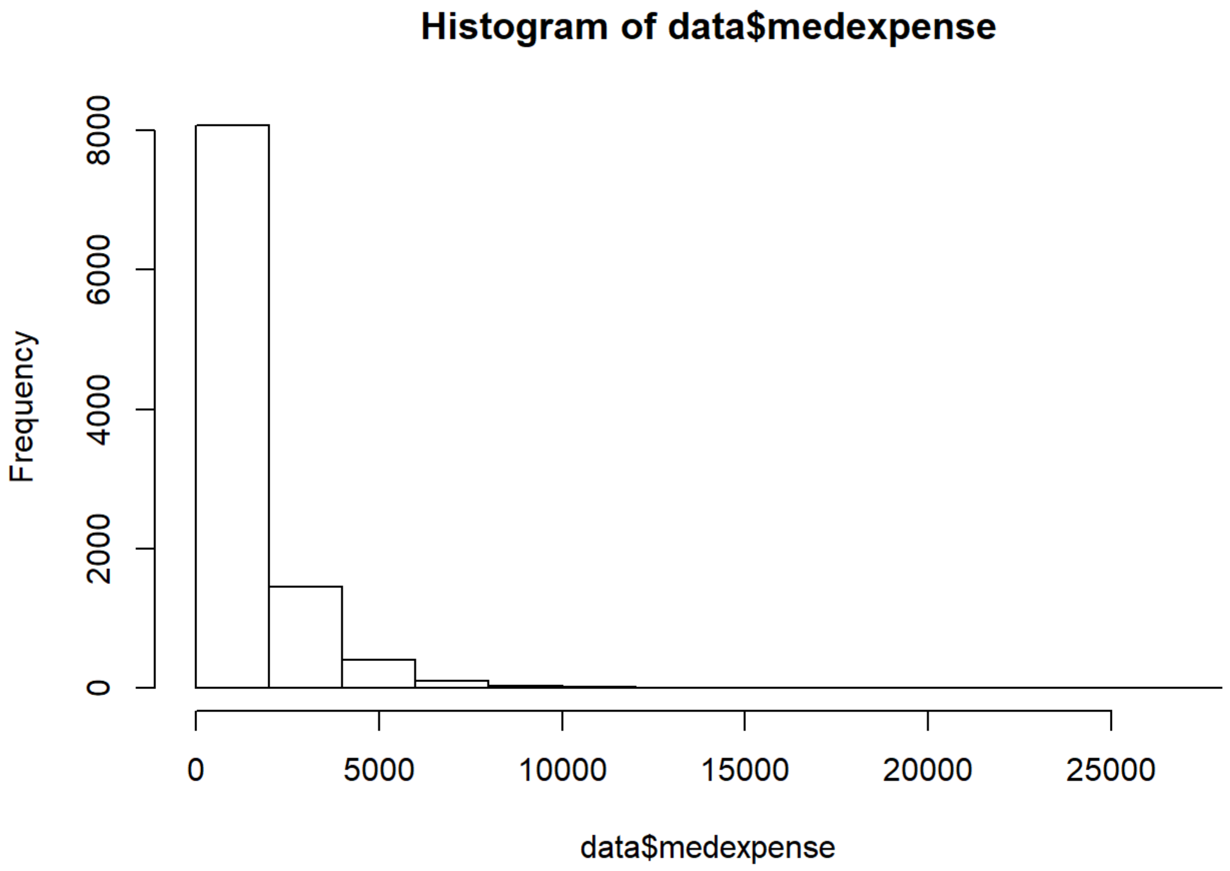
We see that ‘blackhisp’ is highly correlated with ‘black’ and ‘hisp’.
We also notice that ‘healthinsu’ is highly correlated with ‘private’.

```
X = data[, 3:(length(data)-1)] #extract independent variables
corr1 = cor(X) #find correlation
corrplot::corrplot(corr1) #plot correlation plot
```

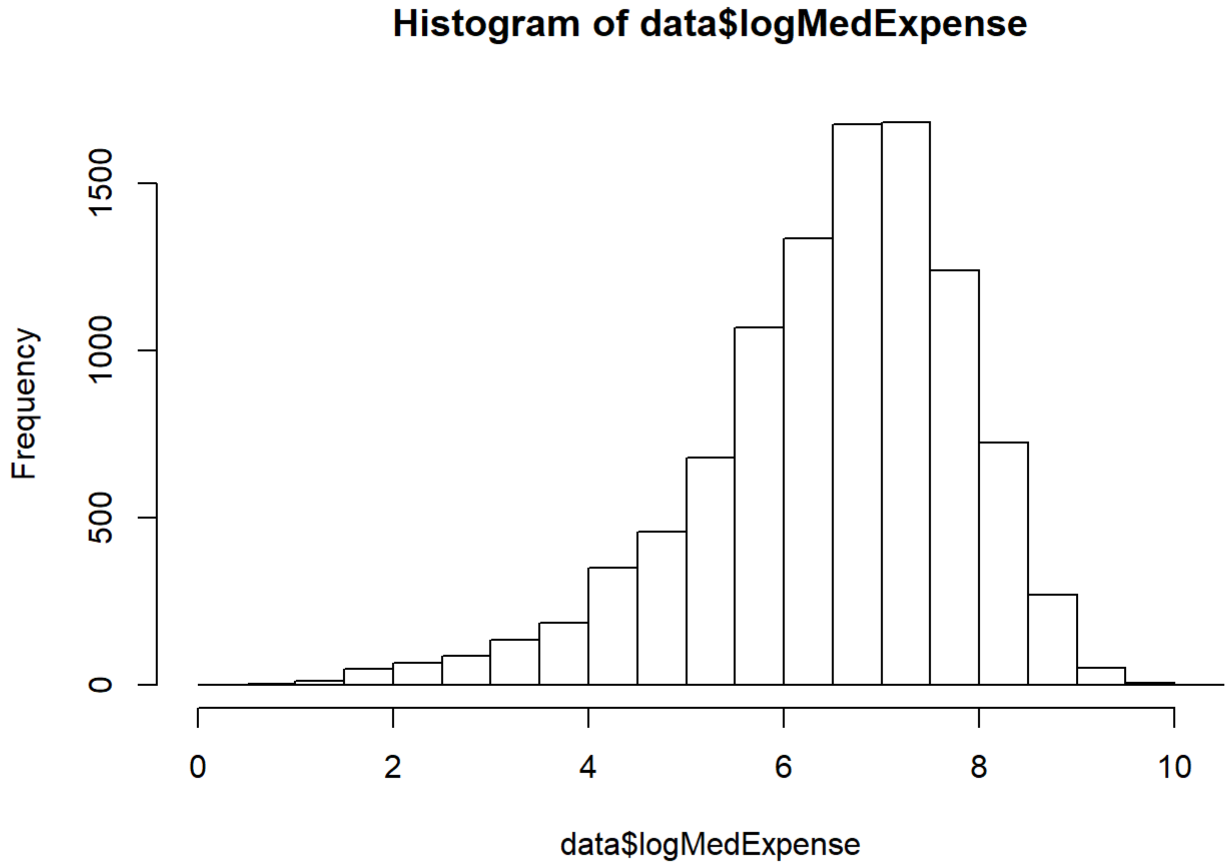


Create factor and dummy variables and log transform medical expenses

```
#Create dummy variables
data$healthinsu = factor(data$healthinsu)
data$female = factor(data$female)
data$marry = factor(data$marry)
data$blackhisp = factor(data$blackhisp)
data$hisp = factor(data$hisp)
data$black = factor(data$black)
data$vegood = factor(data$vegood)
data$good = factor(data$good)
data$fair = factor(data$fair)
data$poor = factor(data$poor)
data$msa = factor(data$msa)
data$private = factor(data$private)
data$priolist = factor(data$priolist)

#log transform medical expenses
hist(data$medexpense)
```



```
data$logMedExpense = log(data$medexpense)
hist(data$logMedExpense)
```



Run regression!

```
#==================== Run regressions ========================

#for res1, we will take all the available attributes
res1 = lm(logMedExpense ~ healthinsu + illnesses + age + female + income + educyr + marry + b
lackhisp + hisp + black + vegood + good + fair + msa + private + priolist, data = data)

#for res2, we remove 'black' & 'hisp' since they are highly correlated with 'blackhisp'
#remove 'private' as it is highly correlated with 'healthinsu'
res2 = lm(logMedExpense ~ healthinsu + illnesses + age + female + income + educyr + marry + b
lackhisp + vegood + good + fair + msa + priolist, data = data)

#for res3 interact 'healthinsu' with 'msa'
res3 = lm(logMedExpense ~ healthinsu*msa + illnesses + age + female + income + educyr + marry
+ blackhisp + vegood + good + fair + priolist, data = data)

#for res4 we interact 'healthinsu' with 'msa', 'illnesses' with 'age', 'illnesses' with 'prio
list' and 'income' with 'educyr'
res4 = lm(logMedExpense ~ healthinsu*msa + illnesses*age + illnesses*priolist + income*educyr
+ female + marry + blackhisp + vegood + good + fair, data = data)

#for our final model we remove 'marry' as it is not significant
res5 = lm(logMedExpense ~ healthinsu*msa + illnesses*age + illnesses*priolist + income*educyr
+ female + blackhisp + vegood + good + fair, data = data)



# ================== Compare Models =====================

#compare model-3 and model-4
AIC(res3, res4) #since AIC score for model 4 is lower, it is better
```

```
##      df      AIC
## res3 16 32630.22
## res4 19 32571.07
```

```
BIC(res3, res4) #since BIC score for model 4 is lower, it is better
```

```
##      df      BIC
## res3 16 32745.73
## res4 19 32708.24
```

```
#compare model-4 and model-5
anova(res4, res5, test = "Chisq") #the p-val shows that keeping 'marry' does not improve the
 model significantly. Therefore, model-5 (without 'marry') is better.
```

```
## Analysis of Variance Table
##
## Model 1: logMedExpense ~ healthinsu * msa + illnesses * age + illnesses *
##     priolist + income * educyr + female + marry + blackhisp +
##     vegood + good + fair
## Model 2: logMedExpense ~ healthinsu * msa + illnesses * age + illnesses *
##     priolist + income * educyr + female + blackhisp + vegood +
##     good + fair
##   Res.Df   RSS Df Sum of Sq Pr(>Chi)
## 1  10071 14853
## 2  10072 14853 -1  -0.18556   0.7228
```

```
#At this point we conclude, model-5 is better than all models presented here.

# ================== Show Results ========================

#show regression results
stargazer(res1, res2, res3, res4, res5,
          title="Regression Results", type="text",
          column.labels=c("Model-1", "Model-2", "Model-3", "Model-4", "Model-5"),
          df=FALSE, digits=3, star.cutoffs = c(0.05,0.01,0.001))
```

```
##
## Regression Results
## ==========================================================================
##                                      Dependent variable:
##                     ------------------------------------------------------
##                                         logMedExpense
##                       Model-1    Model-2    Model-3    Model-4    Model-5
##                         (1)        (2)        (3)        (4)        (5)
## --------------------------------------------------------------------------
## healthinsu1          0.087**    0.080**    0.226***   0.227***   0.226***
##                      (0.032)    (0.026)    (0.051)    (0.051)    (0.051)
##
## illnesses            0.371***   0.371***   0.371***   0.941***   0.941***
##                      (0.010)    (0.010)    (0.010)    (0.112)    (0.112)
##
## age                  -0.005*    -0.005*    -0.005*     0.003      0.003
##                      (0.002)    (0.002)    (0.002)    (0.003)    (0.003)
##
## female1               0.063*     0.063*     0.061*    0.068**    0.071**
##                      (0.026)    (0.026)    (0.026)    (0.026)    (0.025)
##
## income               -0.0001    -0.0001    -0.00003    0.006*     0.006*
##                      (0.001)    (0.001)    (0.001)    (0.003)    (0.003)
##
## educyr                0.007      0.007      0.007     0.014**    0.014**
##                      (0.004)    (0.004)    (0.004)    (0.006)    (0.006)
##
## marry1                -0.012     -0.012     -0.013     -0.010
##                      (0.027)    (0.027)    (0.027)    (0.027)
##
## blackhisp1            -0.117   -0.159***  -0.160***  -0.155***  -0.154***
##                      (0.252)    (0.034)    (0.034)    (0.035)    (0.034)
##
## hisp1                 -0.045
##                      (0.247)
##
## black1                -0.043
##                      (0.249)
##
## vegood1               0.019      0.018      0.017      0.010      0.010
##                      (0.036)    (0.036)    (0.036)    (0.035)    (0.035)
##
## good1                0.101**    0.101**    0.099**    0.091**    0.091**
##                      (0.034)    (0.034)    (0.034)    (0.034)    (0.034)
##
## fair1                0.259***   0.259***   0.257***   0.251***   0.250***
##                      (0.039)    (0.039)    (0.039)    (0.039)    (0.039)
##
## msa1                  -0.047     -0.047     0.018      0.020      0.020
##                      (0.028)    (0.028)    (0.034)    (0.034)    (0.034)
##
## private1              -0.012
##                      (0.032)
##
## priolist1            0.583***   0.584***   0.584***   0.792***   0.792***
##                       (0.039)    (0.039)    (0.038)    (0.049)    (0.049)
##
## healthinsu1:msa1                          -0.193***  -0.193***  -0.193***
##                                            (0.058)    (0.058)    (0.058)
##
## illnesses:age                                         -0.004**   -0.004**
##                                                       (0.001)    (0.001)
##
## illnesses:priolist1                                  -0.287***  -0.287***
##                                                       (0.040)    (0.040)
##
## income:educyr                                         -0.0004*   -0.0004*
##                                                       (0.0002)   (0.0002)
##
## Constant             5.473***   5.471***   5.418***   4.603***   4.584***
##                       (0.167)    (0.166)    (0.167)    (0.263)    (0.257)
##
## --------------------------------------------------------------------------
## Observations          10,089     10,089     10,089     10,089     10,089
## R2                     0.200      0.200      0.201      0.206      0.206
## Adjusted R2            0.199      0.199      0.200      0.205      0.205
## Residual Std. Error    1.219      1.219      1.218      1.214      1.214
## F Statistic         157.758*** 194.205*** 181.292*** 154.059*** 163.694***
## ==========================================================================
## Note:                                         *p<0.05; **p<0.01; ***p<0.001
```

Results

Interpretation of the best model (Model-5):

- **healthinsu**: Having a medical insurance is associated with an increase in medical expenses by 22.6% versus not having a medical insurance.
- **illnesses**: One unit increase in total illnesses in a year is associated with an increase in medical expenses by 94.1%.
- **age**: One year increase in age is associated with an increase in medical expenses by 0.3%.
- **female**: A female patient is associated with 7.1% higer medical expenses than a male.
- **income**: One unit increase in income is associated with an increase in medical expenses by 0.6%.
- **educyr**: One unit increase in years of education is associated with 1.4% increase in medical expenses.
- **blackhisp**: The medical expenses, when the patient’s race is either black or hispanic, is 15.4% lower than when the race is neither black nor hispanic.
- **vegood**: The medical expenses when the patient’s condition is very good is 1% higher than when the condition is poor. However, this result is not significant, since the p-value is greater than 0.05.
- **good**: The medical expenses when the patient’s condition is good is 9.1% higher than when the condition is poor.
- **fair**: The medical expenses when the patient’s condition is fair is 25% higher than when the condition is poor.
- **priolist**: Having a condition listed on priority list is associated with 79.2% increase in medical expenses.
- **healthinsu:msa**: For a person located in an urban area, the impact of having a medical insurance on medical expenses decreases by 19.3% versus when the person is not located in an urban area.
- **illnesses:age**: As age of person increases by one year, the impact of illnesses on medical expenses decreases by 0.4%.
- **illnesses:priolist**: For a person having a condition listed on the priority list, the impact of illnesses on medical expenses decreases by 28.7% versus when the condition is not listed on the priority list.

- **income:educyr**: As years of education increases by one unit, the impact of income on medical expenses decreases by 0.04%.

Based on what metrics did you choose the “best” model?

Model-1 consists of all the attributes present in our dataset (except ssiratio and indid). From this point we start with a backward elimination process. We remove the attributes ‘black’ and ‘hispanic’ since they both were highly correlated with ‘blackhisp’. We also remove the attribute ‘private’ as it was highly correlated with our key independent variable ‘healthinsu’.

We introduce our first interaction term in model-3 as ‘healthinsu:msa’. In model-4 we introduce two more interaction terms as ‘illnesses:age’ and ‘illnesses:priolist’. To check if these additional interaction terms significantly improved our model, we compare model-3 and model-4. Since these two models are not nested, we compare them using AIC and BIC tests. The smaller AIC BIC values for model-4 suggested that model-4 is significantly better than model-3.

At this stage, we have an insignificant attribute ‘marry’ in model-4. We create a new model called model-5 without the ‘marry’ attribute. To test if the presence of attribute ‘marry’ significantly improves our model, we compare model-4 and model-5. Since these are nested models, we use Likelihood ratio test for comparison. The insignificant p-value from this test suggests that presence of ‘marry’ does not significantly improve our model. Therefore, model-5 (without attribute ‘marry’) is better.

Hence, out of all the models we disscussed, model-5 best fits our data!

Do you think the endogeneity of the $HealthIns$ variable still exists? Why or why not?

Yes!‘Healthinsu’ is still endogeneous. There could be an omitted variable (like “previous medical history” or ‘whether patient exercises’) which is correlated with both key independent variable and dependent variable i.e. with ‘healthinsu’ and ‘medExpense’. This omitted variable will cause an omitted variable bias and in turn cause endogineity.
Also, theoretically we would assume that having a health insurance would lower your medical expenses. This assumption is violated by our OLS model, which suggests having health insurance will increase medical expenses. Therefore, we could assume that our model has endogineity.

Question 2

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

a. Use `ivreg()` conduct the 2SLS estimates for your “best” model, while correcting for endogeneity of the $HealthIns$ variable.
b. 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.”

Run Ivreg/2SLS model

```
#run Ivreg/2SLS model
model1 = ivreg(logMedExpense ~ healthinsu + msa + illnesses*age + illnesses*priolist + income
*educyr + female + blackhisp + vegood + good + fair | ssiratio + msa + illnesses*age + illnes
ses*priolist + income*educyr + female + blackhisp + vegood + good + fair, data = data)

summary(model1, diagnostics = TRUE)
```

```
## 
## Call:
## ivreg(formula = logMedExpense ~ healthinsu + msa + illnesses * 
##     age + illnesses * priolist + income * educyr + female + blackhisp + 
##     vegood + good + fair | ssiratio + msa + illnesses * age + 
##     illnesses * priolist + income * educyr + female + blackhisp + 
##     vegood + good + fair, data = data)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -6.8080 -0.7650  0.1163  0.9016  3.9240 
## 
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)    
## (Intercept)          5.3837733  0.3247056  16.580  < 2e-16 ***
## healthinsu1         -1.0796919  0.2509917  -4.302 1.71e-05 ***
## msa1                 0.0483346  0.0365662   1.322 0.186252    
## illnesses            0.9663652  0.1231355   7.848 4.66e-15 ***
## age                 -0.0068929  0.0040971  -1.682 0.092522 .  
## priolist1            0.7693765  0.0535849  14.358  < 2e-16 ***
## income               0.0151147  0.0037366   4.045 5.27e-05 ***
## educyr               0.0426076  0.0085384   4.990 6.14e-07 ***
## female1             -0.0205697  0.0337103  -0.610 0.541750    
## blackhisp1          -0.2087981  0.0395005  -5.286 1.28e-07 ***
## vegood1              0.0205567  0.0388745   0.529 0.596959    
## good1                0.0869105  0.0368439   2.359 0.018349 *  
## fair1                0.2326536  0.0429047   5.423 6.01e-08 ***
## illnesses:age       -0.0047307  0.0015667  -3.019 0.002538 ** 
## illnesses:priolist1 -0.2396184  0.0452051  -5.301 1.18e-07 ***
## income:educyr       -0.0008907  0.0002425  -3.673 0.000241 ***
## 
## Diagnostic tests:
##                    df1   df2 statistic p-value    
## Weak instruments     1 10073    130.58 < 2e-16 ***
## Wu-Hausman           1 10072     25.99 3.5e-07 ***
## Sargan               0    NA        NA      NA    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 1.33 on 10073 degrees of freedom
## Multiple R-Squared: 0.04777, Adjusted R-squared: 0.04635 
## Wald test: 145.6 on 15 and 10073 DF,  p-value: < 2.2e-16
```

Results

Wald Test for our model shows a significant p-value. This means our 2SLS fits our data significantly!
We also notice that residual standard error of 2SLS model is greater than OLS model (1.330 > 1.214). This happens because in 2SLS, stage 2 of regression is based on estimates of stage 1 regression. These subsequent stages of regression causes the standard errors to increase.

```
# =================== Compare Results =========================

#show regression results
stargazer(res5, model1,
          title="Regression Results", type="text",
          df=FALSE, digits=3, star.cutoffs = c(0.05,0.01,0.001))
```

```
##
## Regression Results
## ==================================================
##                           Dependent variable:
##                      ------------------------------
##                              logMedExpense
##                            OLS        instrumental
##                                         variable
##                            (1)            (2)
## --------------------------------------------------
## healthinsu1            0.226***       -1.080***
##                         (0.051)        (0.251)
##
## msa1                     0.020          0.048
##                         (0.034)        (0.037)
##
## illnesses              0.941***        0.966***
##                         (0.112)        (0.123)
##
## age                      0.003          -0.007
##                         (0.003)        (0.004)
##
## priolist1              0.792***        0.769***
##                         (0.049)        (0.054)
##
## income                  0.006*         0.015***
##                         (0.003)        (0.004)
##
## educyr                  0.014**        0.043***
##                         (0.006)        (0.009)
##
## female1                 0.071**         -0.021
##                         (0.025)        (0.034)
##
## blackhisp1             -0.154***      -0.209***
##                         (0.034)        (0.040)
##
## vegood1                  0.010          0.021
##                         (0.035)        (0.039)
##
## good1                   0.091**         0.087*
##                         (0.034)        (0.037)
##
## fair1                  0.250***        0.233***
##                         (0.039)        (0.043)
##
## healthinsu1:msa1       -0.193***
##                         (0.058)
##
## illnesses:age          -0.004**        -0.005**
##                         (0.001)        (0.002)
##
## illnesses:priolist1    -0.287***      -0.240***
##                         (0.040)        (0.045)
##
```

```
## income:educyr          -0.0004*       -0.001***
##                        (0.0002)        (0.0002)
##
## Constant               4.584***        5.384***
##                         (0.257)        (0.325)
##
## -----------------------------------------------------
## Observations             10,089          10,089
## R2                        0.206          0.048
## Adjusted R2               0.205          0.046
## Residual Std. Error      1.214          1.330
## F Statistic           163.694***
## =====================================================
## Note:                *p<0.05; **p<0.01; ***p<0.001
```

Interpretation of the IVreg/2SLS model:

- **healthinsu**: Having a medical insurance is associated with an 108% decrease in medical expenses versus not having a medical insurance. **This is opposite of what OLS model had predicted!**
- **illnesses**: One unit increase in total illnesses in a year is associated with an increase in medical expenses by 96.6%.
- **age**: One year increase in age is associated with an decrease in medical expenses by 0.7%. However this result is not significant as the p-value is greater than 0.05.
- **priolist**: Having a condition listed on priority list is associated with 76.9% increase in medical expenses.
- **income**: One unit increase in income is associated with an increase in medical expenses by 1.5%.
- **educyr**: One unit increase in years of education is associated with 4.3% increase in medical expenses.
- **female**: A female patient is associated with 2.1% lower medical expenses than a male. However this result is not significant as the p-value is greater than 0.05.
- **blackhisp**: The medical expenses, when the patient’s race is either black or hispanic, is 20.9% lower than when the race is neither black nor hispanic.
- **vegood**: The medical expenses when the patient’s condition is very good is 2.1% higher than when the condition is poor. However, this result is not significant, since the p-value is greater than 0.05.
- **good**: The medical expenses when the patient’s condition is good is 8.7% higher than when the condition is poor.
- **fair**: The medical expenses when the patient’s condition is fair is 25% higher than when the condition is poor.
- **illnesses:age**: As age of person increases by one year, the impact of illnesses on medical expenses decreases by 0.5%.
- **illnesses:priolist**: For a person having a condition listed on the priority list, the impact of illnesses on medical expenses decreases by 24% versus when the condition is not listed on the priority list.
- **income:educyr**: As years of education increases by one unit, the impact of income on medical expenses decreases by 0.1%.

Based on our 2SLS model, we can say that having a health insurance decreases our medical expenses! And this is in line with our theoretical assumption!