Homework 1 for Marketing Aanlytics

Regression Analysis

Lunjing Yuan

Due on Monday, January 18, 2020

Feel free to conduct your analysis using this R notebook file. For help with R and R graphics, please check the class notes and the additional notes posted on the class drive.

Please Knit to pdf file, then submit it on Camino by 5:30pm on Monday, Jan. 18. Please name the file in the format DongXiaojing_session1_hw1.pdf before submission.

Part I: Basics of Regression

Follow the steps below:

1. Put the data and this file in a folder, and set it as your working folder through `setwd()`

```
#setwd(" ")
setwd("~/Desktop/winter2021/Marketing Analystics/Lecture 2 Regression")
```

2. Read in the data file `Coffee_inClass.csv`, Run a regression analysis and answer the question "how price influence sales"? You can try different model specificatoin, but only leave the final version of your code here. Make sure you include some dummy variables, and interactions between some dummy with other variables.

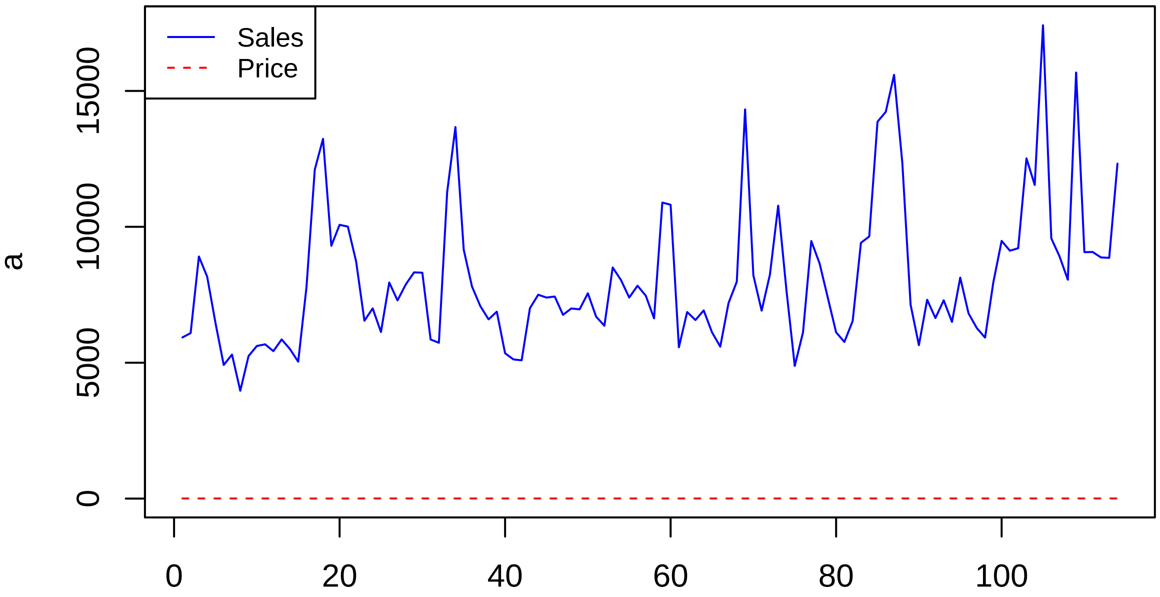
```
#install.packages("ggplot2")
library(ggplot2)
coffee.data<-read.csv("Coffee_inClass.csv")

names(coffee.data)
```

```
## [1] "day"       "dayofweek" "Sales1"    "Price1"    "feat1"     "disp1"
```

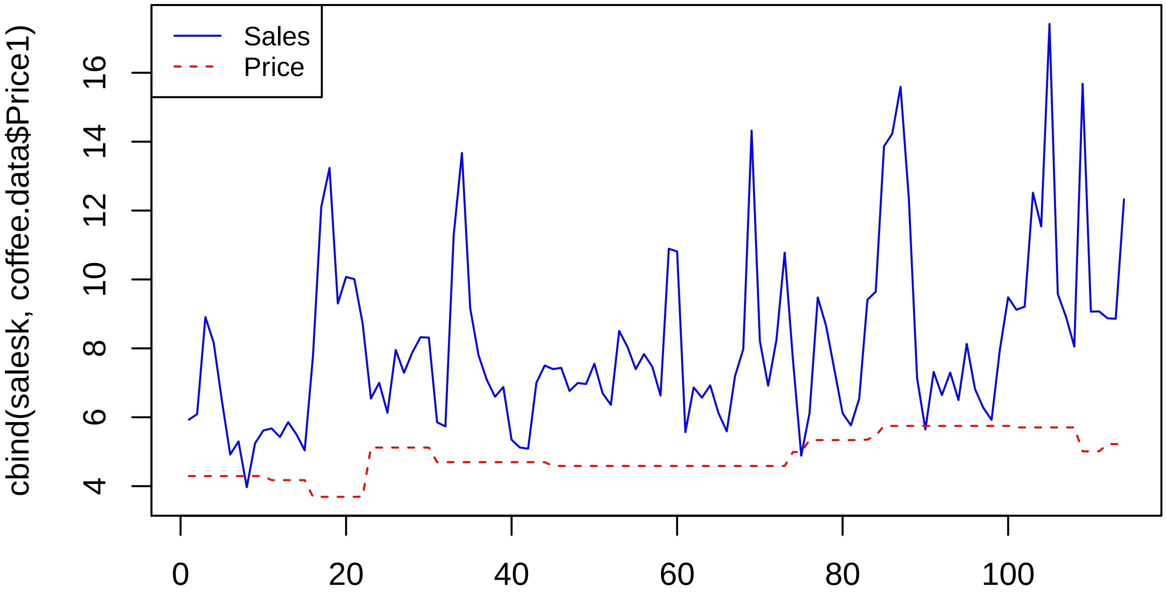
```
#Plot the data
a = cbind(coffee.data$Sales1,coffee.data$Price1)
dim(a)
```

```
## [1] 114   2
```

```
matplot(a,type="l",col=c("blue","red"))
legend('topleft',c("Sales","Price"),lty=1:2,col=c("blue","red"),cex=0.8)
```



```
#We found that the variable Sales1 is in thousands, and the variable Price1 is in units. In order to ma
salesk=coffee.data$Sales1/1000
matplot(cbind(salesk,coffee.data$Price1),type="l",col=c("blue","red"))
legend('topleft',c("Sales","Price"),lty=1:2,col=c("blue","red"),cex=0.8)
```



```
## Final version of regression model

dum_dayofweek <- factor(coffee.data$dayofweek)
salesk=coffee.data$Sales1/1000
model <- lm(salesk~Price1+feat1+dum_dayofweek+feat1*dum_dayofweek,data=coffee.data)
summary(model)
```

```
##
## Call:
## lm(formula = salesk ~ Price1 + feat1 + dum_dayofweek + feat1 *
##     dum_dayofweek, data = coffee.data)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.3107 -1.1673 -0.2434  0.6806  6.3540
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           2.424799   1.832940   1.323  0.18895
## Price1                0.804979   0.336933   2.389  0.01880 *
## feat1                 0.077860   0.024701   3.152  0.00215 **
## dum_dayofweek2        0.758129   0.982772   0.771  0.44231
## dum_dayofweek3        1.756896   0.982870   1.788  0.07695 .
## dum_dayofweek4        0.668164   1.035136   0.645  0.52012
## dum_dayofweek5        0.759577   1.005527   0.755  0.45182
## dum_dayofweek6        0.118542   1.008729   0.118  0.90669
## dum_dayofweek7        0.357953   1.038207   0.345  0.73100
## feat1:dum_dayofweek2 -0.007708   0.044805  -0.172  0.86376
## feat1:dum_dayofweek3 -0.116792   0.037966  -3.076  0.00272 **
## feat1:dum_dayofweek4 -0.031117   0.029661  -1.049  0.29671
## feat1:dum_dayofweek5  0.013195   0.032634   0.404  0.68686
## feat1:dum_dayofweek6 -0.008638   0.033980  -0.254  0.79987
## feat1:dum_dayofweek7 -0.020065   0.030525  -0.657  0.51251
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.067 on 98 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.4232, Adjusted R-squared:  0.3408
## F-statistic: 5.135 on 14 and 98 DF,  p-value: 3.865e-07
```

As we can see from the result, for every one dollar change of price1 the expected number of sales increases by 0.804979 on average holding all other variables constant. In real world, it is hard to measure sales based on price. Therefore, we need to make assumptions:

Null hypothesis: price has no influence on sales

Alternative hypothesis: price has influence on sales

From the result above, Price1(0.018) < significance level(0.05), which means that we reject the null hypothesis. Hence, price has influence on sales.

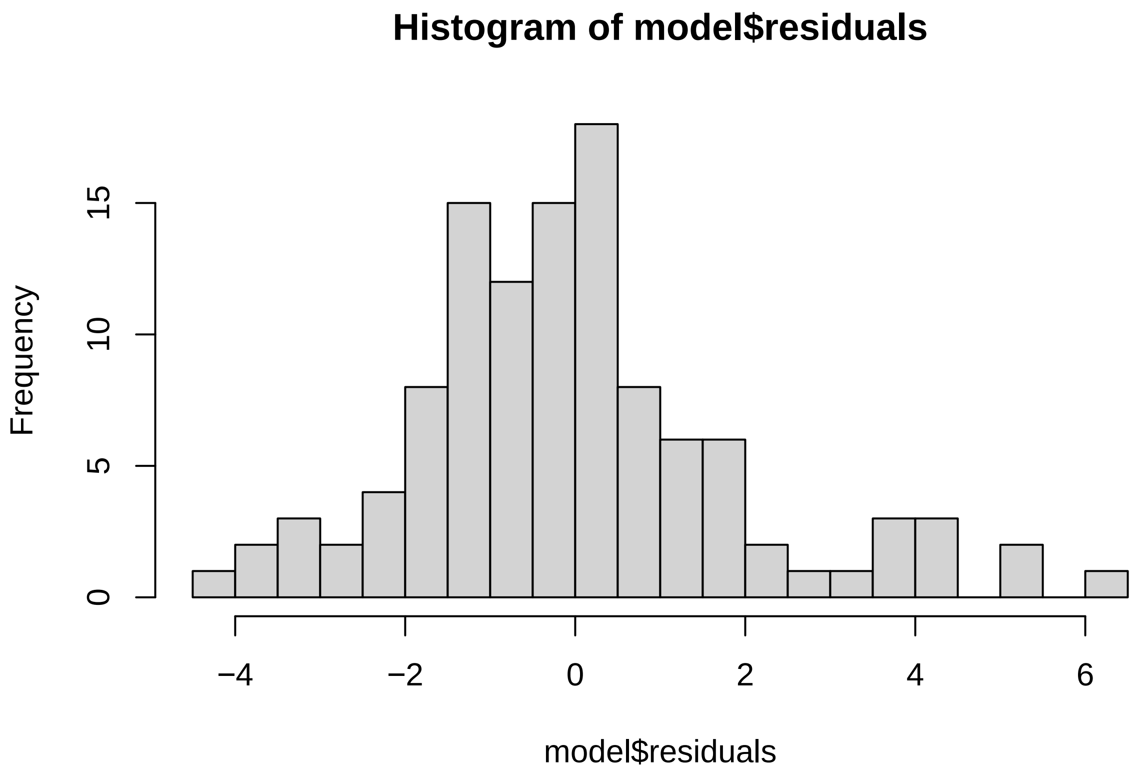
2. List all the control variables (including dummy variables, and interactions) included in the model. Explain for each control variable, why it needs to be included? ###### ## Control variable: Price1 means the price of coffee. As we can see from the dataset, the price is various on different days. That is to say, promotions, membership, discounts and son on may affect the price. Price1 is our interest of variable. feat1 is the featured ads. According to our dataset, cafe shop use feat1 on different days. This specific advertising material could help increase sales.

Dummy variable: dayofweek is a categorical variable with fixed set of values, from Monday to Sunday. Sales may be different during weekdays and at weekends.

Interaction variable: feat1 * dayofweek, indicates that the consequences of feature on sales may be different if the feature is on different days of the week.

3. Plot the residuals, and comment on the residules, are they ideal? Any concerns?

```
#Plotting histogram for residuals
hist(model$residuals,20)
```



This model doesn't work or needs more data because the plot is clearly not a symmetrical pattern.

4. How do you interpret each of the parameter estimates? Make sure your interpretation of each estimates include the values of the estimates, the standard error, the t-statistics and the p-value. Be careful with the dummy variables and the interaction variables.

```
#The variable Sales1 is in thousands, and the variable Price1 is in units. In order to see them togethe

salesk=coffee.data$Sales1/1000

modelx <- lm(salesk~Price1+feat1+dum_dayofweek+feat1*dum_dayofweek,data=coffee.data)
summary(modelx)
```

```
##
## Call:
## lm(formula = salesk ~ Price1 + feat1 + dum_dayofweek + feat1 *
##     dum_dayofweek, data = coffee.data)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.3107 -1.1673 -0.2434  0.6806  6.3540
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           2.424799   1.832940   1.323  0.18895
## Price1                0.804979   0.336933   2.389  0.01880 *
## feat1                 0.077860   0.024701   3.152  0.00215 **
## dum_dayofweek2        0.758129   0.982772   0.771  0.44231
## dum_dayofweek3        1.756896   0.982870   1.788  0.07695 .
## dum_dayofweek4        0.668164   1.035136   0.645  0.52012
## dum_dayofweek5        0.759577   1.005527   0.755  0.45182
## dum_dayofweek6        0.118542   1.008729   0.118  0.90669
## dum_dayofweek7        0.357953   1.038207   0.345  0.73100
## feat1:dum_dayofweek2 -0.007708   0.044805  -0.172  0.86376
## feat1:dum_dayofweek3 -0.116792   0.037966  -3.076  0.00272 **
## feat1:dum_dayofweek4 -0.031117   0.029661  -1.049  0.29671
## feat1:dum_dayofweek5  0.013195   0.032634   0.404  0.68686
## feat1:dum_dayofweek6 -0.008638   0.033980  -0.254  0.79987
## feat1:dum_dayofweek7 -0.020065   0.030525  -0.657  0.51251
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.067 on 98 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.4232, Adjusted R-squared:  0.3408
## F-statistic: 5.135 on 14 and 98 DF,  p-value: 3.865e-07
```

Price1: the estimate is 0.804979, which means for one dollar change of Price1 the expected number of sales increases by 0.804979 on average holding all other variables constant. Standard error is 0.336933, which indicates the data is certain about this parameter as SE value is very small.The estimate is considered positive because t-statistics(2.389) $>$ 2.0. Beisdes, p-value(0.01880) is smaller than significance level 0.05 so price1 is statistically significant.

feat1:the estimate 0.077860 means for one unit change of feat1 the expected number of sales increases by 0.077860 on average holding all other variables constant. Standard error(0.02470) shows the data is certain about this this parameter. The estimate is considered positive because t-statistics(3.152) $>$ 2.0. P-value(0.00215) is smaller than significance level 0.05 so feat1 is statistically significant.

Dummy variables dum_dayofweek2 to dum_dayofweek7. All estimates have positive relationship with sales. T-values of all dummy variables are bewteen -2.0 to 2.0 , hence, estimates are considered statistically zero. Standard error values show that the data is unceratin about this parameter.Lastly, all the p values are greater than significance level hence, they are statistically insignificant. Besides, we remove dum_dayofweek1 (Monday) dummy variable and use dum_dayofweek starts from Tuesday. Take Thursday sales as example, we know that estimated sales on Thursday are 668.164 compared to estimated sales on Monday holding all other variables constant.

Interaction variables feat1 and dum_dayofweek2 - dum_dayofweek7. Estimates have negative relationship with sales. We copuld say for one unit change of feat1*dum_dayofweek2 (Tuesday) the expected number of sales decreases by -7.708 on average holding all other variables constant. T-values of feat1:dum_dayofweek3 is t$-$stat < -2.0 indicating the estimate considered to be negative and p-value 0.00272 $<$ 0.05. Therefore, feat1:dum_dayofweek3 is statistically significant. All other t-values of interaction variables are greater than -2.0 means that the estimate considered to be statistically zero. Standard error values are greater than estimated coefficient means data is very uncertain about this parameter.

5. In utilizing the dummy variables indicating the day of week, the above model has left one of the day-of-week dummy variable out. Now change the specification by leaving out a different day-of-week dummy variable (for example instead of leaving out the Monday dummy, now include the Monday dummy but leave out the Tuesday (or any other day) dummy). Please explain the changes in the estimates, standard errors of all the estimate.

```r
## Removing Tuesday dum_dayofweek2
remove = relevel(dum_dayofweek,ref = "2")
#Run regression with releveled variable
modely <- lm(salesk~Price1+feat1+dum_dayofweek+feat1*remove, data = coffee.data)
summary(modely)
```

```
##
## Call:
## lm(formula = salesk ~ Price1 + feat1 + dum_dayofweek + feat1 *
##     remove, data = coffee.data)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.3107 -1.1673 -0.2434  0.6806  6.3540
##
## Coefficients: (6 not defined because of singularities)
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     2.4247988  1.8329404   1.323   0.1889
## Price1          0.8049791  0.3369329   2.389   0.0188 *
## feat1           0.0701521  0.0373381   1.879   0.0632 .
## dum_dayofweek2  0.7581294  0.9827717   0.771   0.4423
## dum_dayofweek3  1.7568962  0.9828702   1.788   0.0769 .
## dum_dayofweek4  0.6681643  1.0351364   0.645   0.5201
## dum_dayofweek5  0.7595774  1.0055267   0.755   0.4518
## dum_dayofweek6  0.1185415  1.0087291   0.118   0.9067
## dum_dayofweek7  0.3579535  1.0382070   0.345   0.7310
## remove1                NA         NA      NA       NA
## remove3                NA         NA      NA       NA
## remove4                NA         NA      NA       NA
## remove5                NA         NA      NA       NA
## remove6                NA         NA      NA       NA
## remove7                NA         NA      NA       NA
## feat1:remove1   0.0077082  0.0448049   0.172   0.8638
## feat1:remove3  -0.1090838  0.0472988  -2.306   0.0232 *
## feat1:remove4  -0.0234087  0.0406862  -0.575   0.5664
## feat1:remove5   0.0209027  0.0428635   0.488   0.6269
## feat1:remove6  -0.0009296  0.0440086  -0.021   0.9832
## feat1:remove7  -0.0123570  0.0416390  -0.297   0.7673
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.067 on 98 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.4232, Adjusted R-squared:  0.3408
## F-statistic: 5.135 on 14 and 98 DF,  p-value: 3.865e-07
```

Based on our regression model, we find changes in estimates of control variable feat1, interaction variables feat1:remove1,feat1:remove2,feat1:remove4,feat1:remove5,feat1:re

feat1:the estimate 0.0701521 means for every one unit change of feat1 the expected number of sales increases by 0.0701521 on average holding all other variables constant. Standard error(0.0373381) shows the data is certain about this this parameter. The estimate is considered to be statistically zero because -2.0 < t-statistics(1.879) < 2.0. P-value(0.0632) is greater than significance level 0.05 so feat1 is not statistically significant.

As we can see from summary, there is a positive relationship with sales for interaction variables. For every one unit change of feat1:remove1 (Sales on Monday) the expected number of sales increases by 7.7082 on average holding all other variables constant.

T-values of interaction variables feat1:remove1, feat1:remove4, feat1:remove5, feat1:remove6, feat1:remove7 are between -2.0 and 2.0 indicating the estimate considered to be statistically zero.T-values of interaction variables feat1:remove3 is smaller than -2.0 indicating the estimate considered to be negative. Standard error values are less than estimated coefficient indicating data is very certain about this parameter.

Part II Endogeneity and 2SLS

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.

```r
health.data<-read.csv("health_inclass.csv")
attach(health.data)
names(health.data)
```

```
##  [1] "indid"      "medexpense" "healthinsu" "illnesses"  "age"
##  [6] "female"     "income"     "ssiratio"   "educyr"     "marry"
## [11] "blackhisp"  "hisp"       "black"      "vegood"     "good"
## [16] "fair"       "poor"       "msa"        "private"    "priolist"
```

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

```r
Y1 <- log(medexpense)
logincome = log(income)

lm1<- lm(Y1~healthinsu+illnesses)
summary(lm1)
```

```
##
## Call:
## lm(formula = Y1 ~ healthinsu + illnesses)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.2552 -0.6755  0.1487  0.8545  3.7622
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 5.631625   0.023651  238.12  < 2e-16 ***
## healthinsu  0.086209   0.025356    3.40 0.000677 ***
## illnesses   0.438912   0.009531   46.05  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.238 on 10086 degrees of freedom
## Multiple R-squared:  0.1746, Adjusted R-squared:  0.1744
## F-statistic:  1067 on 2 and 10086 DF,  p-value: < 2.2e-16
```

```
lm2<- lm(Y1~healthinsu+illnesses+age)
summary(lm2)
```

```
##
## Call:
## lm(formula = Y1 ~ healthinsu + illnesses + age)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.2701 -0.6749  0.1475  0.8537  3.7787
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.841905   0.142575  40.974  < 2e-16 ***
## healthinsu   0.080284   0.025662   3.129  0.00176 **
## illnesses    0.440036   0.009560  46.030  < 2e-16 ***
## age         -0.002800   0.001872  -1.496  0.13479
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.238 on 10085 degrees of freedom
## Multiple R-squared:  0.1748, Adjusted R-squared:  0.1745
## F-statistic: 711.9 on 3 and 10085 DF,  p-value: < 2.2e-16
```

```
lm3<- lm(Y1~healthinsu+illnesses+age+log(income))
summary(lm3)
```

```
##
## Call:
## lm(formula = Y1 ~ healthinsu + illnesses + age + log(income))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.2793 -0.6768  0.1472  0.8517  3.7803
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.780127   0.150891  38.307  < 2e-16 ***
## healthinsu   0.074959   0.026012   2.882  0.00396 **
## illnesses    0.440653   0.009572  46.035  < 2e-16 ***
## age         -0.002595   0.001879  -1.381  0.16735
## log(income)  0.017236   0.013787   1.250  0.21124
## ---
## 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:  0.1749, Adjusted R-squared:  0.1746
## F-statistic: 534.4 on 4 and 10084 DF,  p-value: < 2.2e-16
```

(1) Based on what metrics did you choose the "best" model?

R Square. A higher R Square value indicates the model fits the data better.

F-statistics evaluates whether the model as a whole is actually necessary.

(2) Do you think the endogeneity of the *HealthIns* variable still exists? Why or why not?

Endogeneity still exists. The problem here is that people's decision on whether to have insurance is not random. It is highly possible that people with health insurance would choose more expensive treatments. Therefore, without running the statistical model, we suspect corr(error term, HealthIns) != 0. Hence, HealthIns is endogenous.

2. Suppose the *HealthIns* is 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.

I ran 3 models here: Simple OLS, 2SLS using ivreg(), 2SLS using two lm(). The last one, Model 3 using ivreg().

```r
health.data = read.csv("health_inclass.csv", header = TRUE)
attach(health.data)
```

```
## The following objects are masked from health.data (pos = 3):
##
##     age, black, blackhisp, educyr, fair, female, good, healthinsu,
##     hisp, illnesses, income, indid, marry, medexpense, msa, poor,
##     priolist, private, ssiratio, vegood
```

```r
Y1 <- log(medexpense)
Y2 <- healthinsu
X1 <- cbind(illnesses, age, logincome)
X2 <- cbind(ssiratio)

#model1
model1 <- lm(Y1 ~ Y2 + X1)
summary(model1)
```

```
##
## Call:
## lm(formula = Y1 ~ Y2 + X1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.2793 -0.6768  0.1472  0.8517  3.7803
##
```

```
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.780127   0.150891  38.307  < 2e-16 ***
## Y2           0.074959   0.026012   2.882  0.00396 **
## X1illnesses  0.440653   0.009572  46.035  < 2e-16 ***
## X1age       -0.002595   0.001879  -1.381  0.16735
## X1logincome  0.017236   0.013787   1.250  0.21124
## ---
## 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:  0.1749, Adjusted R-squared:  0.1746
## F-statistic: 534.4 on 4 and 10084 DF,  p-value: < 2.2e-16
```

```
#model2
# 2SLS estimation (details)
olsreg1 <- lm (Y2 ~ X1 + X2)
summary(olsreg1)
```

```
##
## Call:
## lm(formula = Y2 ~ X1 + X2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.6817 -0.3882 -0.2413  0.5167  2.5921
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.9591576  0.0568776  16.864  < 2e-16 ***
## X1illnesses  0.0113510  0.0036336   3.124  0.00179 **
## X1age       -0.0085302  0.0007125 -11.973  < 2e-16 ***
## X1logincome  0.0544246  0.0056429   9.645  < 2e-16 ***
## X2          -0.1997539  0.0141579 -14.109  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4691 on 10084 degrees of freedom
## Multiple R-squared:  0.06839,    Adjusted R-squared:  0.06803
## F-statistic: 185.1 on 4 and 10084 DF,  p-value: < 2.2e-16
```

```
Y2hat <- fitted(olsreg1)
model2 <- lm(Y1 ~ Y2hat + X1)
summary(model2)
```

```
##
## Call:
## lm(formula = Y1 ~ Y2hat + X1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.2923 -0.6683  0.1525  0.8507  3.6881
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.589839   0.221021  29.815  < 2e-16 ***
## Y2hat       -0.852201   0.186843  -4.561 5.15e-06 ***
## X1illnesses  0.448512   0.009694  46.267  < 2e-16 ***
## X1age       -0.011797   0.002627  -4.492 7.15e-06 ***
## X1logincome  0.097693   0.021157   4.617 3.93e-06 ***
## ---
## 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:  0.1759, Adjusted R-squared:  0.1756
## F-statistic: 538.1 on 4 and 10084 DF,  p-value: < 2.2e-16
```

```
#model3
library(AER)
```

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

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

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

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

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

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

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

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

```
# 2SLS estimation
model3 <- ivreg(Y1 ~ Y2 + X1 | X1 + X2)
summary(model3)
```

```
##
## Call:
## ivreg(formula = Y1 ~ Y2 + X1 | X1 + X2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.7141 -0.7468  0.1288  0.8907  4.0895
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.589839   0.234676  28.081  < 2e-16 ***
## Y2          -0.852201   0.198386  -4.296 1.76e-05 ***
## X1illnesses  0.448512   0.010293  43.575  < 2e-16 ***
## X1age       -0.011797   0.002789  -4.230 2.36e-05 ***
## X1logincome  0.097693   0.022464   4.349 1.38e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.313 on 10084 degrees of freedom
## Multiple R-Squared: 0.07094, Adjusted R-squared: 0.07058
## Wald test: 477.3 on 4 and 10084 DF,  p-value: < 2.2e-16
```

(2) Compare the results from this model with those from the simple OLS approach, in terms of model fit, and your answers to the question "whether having health insurance leads to higher or lower medical expnses."

```
#Comparison of the Parameter Estimates:
ests=cbind(model1$coefficients,model3$coefficients,model2$coefficients)
colnames(ests) = c('OLS', '2SLS-ivreg', '2SLS-2regressions')
ests
```

```
##                      OLS  2SLS-ivreg 2SLS-2regressions
## (Intercept)   5.78012658  6.58983876        6.58983876
## Y2            0.07495950 -0.85220101       -0.85220101
## X1illnesses   0.44065296  0.44851229        0.44851229
## X1age        -0.00259457 -0.01179746       -0.01179746
## X1logincome   0.01723634  0.09769285        0.09769285
```

By comparing the the Parameter Estimates in different models, we noticed the Y2 (HealthIns) estimates have a big difference in those models. Since we corrected the endogeneity, we won't suggest people to skip insurance if they want to save medical cost.

(3) Compare the results in both estimates and the standard errors. The estimates for the endogeneous variables are quite different whether endogeneity is controlled or not.

```
#Comparison of the Standard Errors:
stderrs <- cbind(summary(model1)$coefficients[,2],
                 summary(model3)$coefficients[,2],
                 summary(model2)$coefficients[,2])
colnames(stderrs) = c('OLS', '2SLS-ivreg', '2SLS-2regressions')
stderrs
```

```
##                     OLS  2SLS-ivreg 2SLS-2regressions
## (Intercept) 0.150891039 0.234676090       0.221021221
## Y2          0.026012427 0.198386020       0.186842726
## X1illnesses 0.009572120 0.010292818       0.009693920
## X1age       0.001878955 0.002788866       0.002626593
## X1logincome 0.013786540 0.022464356       0.021157244
```

Given that these are predicted values, the uncertainty in the estimates from the first step regression will lead to uncertainty in these predicted values. That is to say, the standard error for the endogenous variable in the 2SLS-2regressions column is wrong. By comparing the first and second columns, we found that although the standard error for the endogenous variable are very different (0.0260124 in the first column (OLS), vs. 0.198386 in the second column (ivreg)). When we correct for endogeneity, it ensures the parameter estimates to be unbiased and correct, but it introduces more uncertainty into the model estimates.