Case Project

Martin Salvo

Monday, December 21, 2015

Introduction

This is an **R** Markdown document that tries to make reproducible my answers for Test1 is the Erasmus MOOC. The code can be run in **R** and have an easy access to the reasoning I have done.

Loading Data Analysis

```
library(astsa)              # then load it (has to be done at the start of
each session)

## Warning: package 'astsa' was built under R version 3.2.3

library(zoo)

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

library(xts)
library(tseries)

## Warning: package 'tseries' was built under R version 3.2.2

library(devtools)

## Warning: package 'devtools' was built under R version 3.2.3

## WARNING: Rtools is required to build R packages, but is not currently
installed.
##
## Please download and install Rtools 3.3 from http://cran.r-
project.org/bin/windows/Rtools/ and then run find_rtools().

library(stats)

setwd("C:/Users/samsung/Documents/Data_Science/Econometrics")
DataTest <- read.table("CaseData.txt", header=TRUE)

model1 <-lm(sell ~ lot + bdms + fb + sty + drv + rec + ffin + ghw + ca +
gar + reg, data=DataTest)
summary(model1)

##
## Call:
## lm(formula = sell ~ lot + bdms + fb + sty + drv + rec + ffin +
##     ghw + ca + gar + reg, data = DataTest)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -41389  -9307   -591   7353  74875
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -4038.3504  3409.4713  -1.184 0.236762
## lot             3.5463     0.3503  10.124  < 2e-16 ***
## bdms         1832.0035  1047.0002   1.750 0.080733 .
## fb          14335.5585  1489.9209   9.622  < 2e-16 ***
## sty          6556.9457   925.2899   7.086 4.37e-12 ***
## drv          6687.7789  2045.2458   3.270 0.001145 **
## rec          4511.2838  1899.9577   2.374 0.017929 *
## ffin         5452.3855  1588.0239   3.433 0.000642 ***
## ghw         12831.4063  3217.5971   3.988 7.60e-05 ***
## ca          12632.8904  1555.0211   8.124 3.15e-15 ***
## gar          4244.8290   840.5442   5.050 6.07e-07 ***
## reg          9369.5132  1669.0907   5.614 3.19e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15420 on 534 degrees of freedom
## Multiple R-squared:  0.6731, Adjusted R-squared:  0.6664
## F-statistic: 99.97 on 11 and 534 DF,  p-value: < 2.2e-16
```

Answer A

The first model includes all variables, where number of bedrooms (bdms) and having a recreational room (rec) does not seem to be significant. R squared seems to be relatively high at a 67% value. We should re-run this model dropping this variables.

```
##
## Call:
## lm(formula = sell ~ lot + fb + sty + drv + ffin + ghw + ca +
##     gar + reg, data = DataTest)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -42084  -8987   -696   7497  74618
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1123.144   2747.409  -0.409  0.68285
## lot             3.666      0.350  10.476  < 2e-16 ***
## fb          15072.868   1458.857  10.332  < 2e-16 ***
## sty          7241.264    869.559   8.328 6.93e-16 ***
## drv          6428.565   2041.585   3.149  0.00173 ** 
## ffin         7134.099   1481.246   4.816 1.91e-06 ***
## ghw         12954.080   3236.414   4.003 7.15e-05 ***
## ca          12875.657   1560.069   8.253 1.20e-15 ***
## gar          4265.862    842.377   5.064 5.65e-07 ***
## reg          9595.888   1677.353   5.721 1.76e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15520 on 536 degrees of freedom
## Multiple R-squared:  0.6679, Adjusted R-squared:  0.6623
## F-statistic: 119.8 on 9 and 536 DF,  p-value: < 2.2e-16
```

Now all the regressors are significant at a 99% level of confidence.

Answer B

```
sell_log <- log(DataTest$sell, base = exp(1))

DataTest <- cbind(DataTest, sell_log)

model3 <-lm(sell_log ~ lot + bdms + fb + sty + drv + rec + ffin + ghw +
ca + gar + reg, data=DataTest)
summary(model3)
```

```
##
## Call:
## lm(formula = sell_log ~ lot + bdms + fb + sty + drv + rec + ffin +
##     ghw + ca + gar + reg, data = DataTest)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.67865 -0.12211  0.01666  0.12868  0.67737
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.003e+01  4.724e-02 212.210  < 2e-16 ***
## lot         5.057e-05  4.854e-06  10.418  < 2e-16 ***
## bdms        3.402e-02  1.451e-02   2.345  0.01939 *
## fb          1.678e-01  2.065e-02   8.126 3.10e-15 ***
## sty         9.227e-02  1.282e-02   7.197 2.10e-12 ***
## drv         1.307e-01  2.834e-02   4.610 5.04e-06 ***
## rec         7.352e-02  2.633e-02   2.792  0.00542 **
## ffin        9.940e-02  2.200e-02   4.517 7.72e-06 ***
## ghw         1.784e-01  4.458e-02   4.000 7.22e-05 ***
## ca          1.780e-01  2.155e-02   8.262 1.14e-15 ***
## gar         5.076e-02  1.165e-02   4.358 1.58e-05 ***
## reg         1.271e-01  2.313e-02   5.496 6.02e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2137 on 534 degrees of freedom
## Multiple R-squared:  0.6766, Adjusted R-squared:  0.6699
## F-statistic: 101.6 on 11 and 534 DF,  p-value: < 2.2e-16
```

Now all the variables are significant at a 95% level of confidence. R squared is now at a 67% value. The same value obtained in model 1.

Answer C

```
lot_log <- log(DataTest$lot, base = exp(1))
DataTest <- cbind(DataTest, lot_log)

Lot_Test <- adf.test(DataTest$lot)$statistic
```

```
## Warning in adf.test(DataTest$lot): p-value smaller than printed p-
value
```

```
LogLot_Test <- adf.test(DataTest$lot_log)$statistic
```

```
## Warning in adf.test(DataTest$lot_log): p-value smaller than printed p-
value
```

Both series lot and log lot tests are stationary because H0 is rejected at a 95% level of confidence. We can reject both with -4.8351913 and -4.8719404 compared to the t-statistic of -3.41.

Answer D and E

```
library(phia)
```

```
## Warning: package 'phia' was built under R version 3.2.3
```

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

```
## Warning: package 'car' was built under R version 3.2.2
```

```
model4 <-lm(sell_log ~ lot_log + bdms + fb + sty + drv + rec + ffin + ghw
+ ca + gar + reg, data=DataTest)
summary(model4)
```

```
##
## Call:
## lm(formula = sell_log ~ lot_log + bdms + fb + sty + drv + rec +
##     ffin + ghw + ca + gar + reg, data = DataTest)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.68355 -0.12247  0.00802  0.12780  0.67564
##
## Coefficients:
```

```
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  7.74509    0.21634  35.801  < 2e-16 ***
## lot_log      0.30313    0.02669  11.356  < 2e-16 ***
## bdms         0.03440    0.01427   2.410 0.016294 *  
## fb           0.16576    0.02033   8.154 2.52e-15 ***
## sty          0.09169    0.01261   7.268 1.30e-12 ***
## drv          0.11020    0.02823   3.904 0.000107 ***
## rec          0.05797    0.02605   2.225 0.026482 *  
## ffin         0.10449    0.02169   4.817 1.90e-06 ***
## ghw          0.17902    0.04389   4.079 5.22e-05 ***
## ca           0.16642    0.02134   7.799 3.29e-14 ***
## gar          0.04795    0.01148   4.178 3.43e-05 ***
## reg          0.13185    0.02267   5.816 1.04e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.2104 on 534 degrees of freedom
## Multiple R-squared:  0.6865, Adjusted R-squared:  0.6801 
## F-statistic: 106.3 on 11 and 534 DF,  p-value: < 2.2e-16
```

```
anova(model4)
```

```
## Analysis of Variance Table
## 
## Response: sell_log
##            Df  Sum Sq Mean Sq  F value    Pr(>F)    
## lot_log     1 25.3677 25.3677 573.0690 < 2.2e-16 ***
## bdms        1  6.1629  6.1629 139.2224 < 2.2e-16 ***
## fb          1  6.4589  6.4589 145.9085 < 2.2e-16 ***
## sty         1  3.3168  3.3168  74.9270 < 2.2e-16 ***
## drv         1  1.4804  1.4804  33.4434 1.250e-08 ***
## rec         1  1.3491  1.3491  30.4761 5.282e-08 ***
## ffin        1  1.9223  1.9223  43.4249 1.058e-10 ***
## ghw         1  0.3778  0.3778   8.5343  0.003633 ** 
## ca          1  3.0755  3.0755  69.4763 6.586e-16 ***
## gar         1  0.7661  0.7661  17.3077 3.704e-05 ***
## reg         1  1.4975  1.4975  33.8294 1.037e-08 ***
## Residuals 534 23.6383  0.0443                       
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

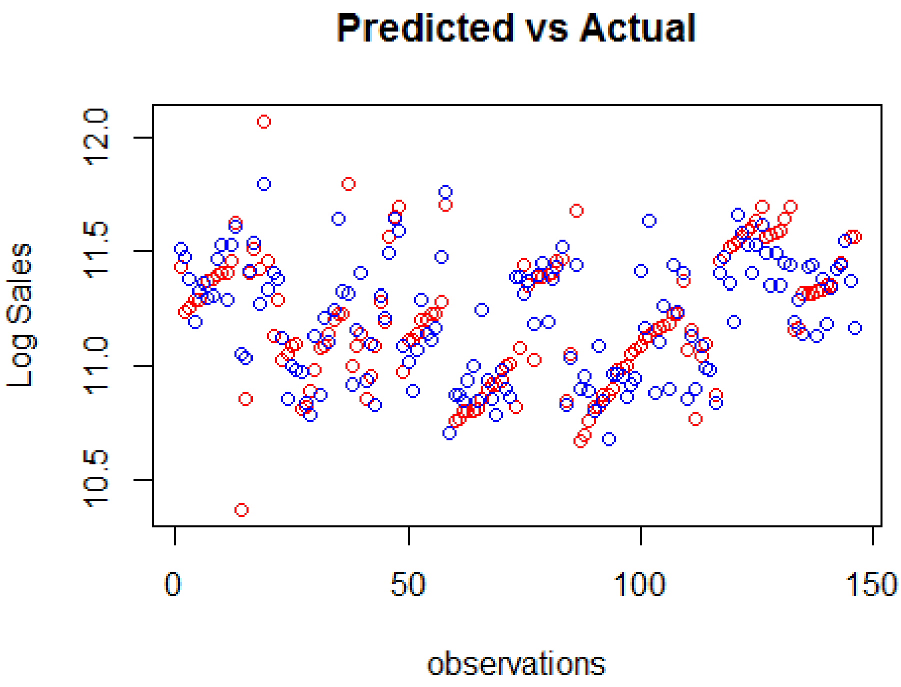
The F-statistic helps to reject the NULL hypothesis so the multiple regression differs from the trivial solution.

Answer G

Condition on the house may be missing and that is why it is important to include it. Using variables as air conditioning may be biased because newer houses are supposed to be in better conditions but this will understimate the price of older houses that are in good condition. According to my view it will underestimate the price of houses.

Answer H

```
train <- DataTest[1:400,]
test <- DataTest[401:546,]

model5 <-lm(sell_log ~ lot_log + bdms + fb + sty + drv + rec + ffin + ghw
+ ca + gar + reg, data=train)

fitted.results <-
predict(model5,newdata=subset(test,select=c(15,4,5,6,7,8,9,10,11,12,13)),
type='response')

observ <- seq(1:146)

Final_Data  <- cbind(observ ,DataTest$sell_log[401:546],fitted.results)
colnames(Final_Data) <- c("observ", "Values", "Predicted")
Final_Data  <- as.data.frame(Final_Data)

plot(Final_Data$observ, Final_Data$Values, type="pcy", col="red", main =
"Predicted vs Actual", ylab= "Log Sales", xlab= "observations")

## Warning in plot.xy(xy, type, ...): gráfico de tipo 'pcy' va a ser
truncado
## al primer carácter

points(Final_Data$observ, Final_Data$Predicted, col="blue")
```



```
RMSE <- mean((Final_Data$Predicted - Final_Data$Values)^2)
MAE <- mean(abs(Final_Data$Predicted - Final_Data$Values))
SFE <- sum(abs(Final_Data$Predicted - Final_Data$Values))

Variance <- var(Final_Data$Values)
```

The model has a decent predictive power as seen in the chart, log variance in somewhat lower than absolute mean error (MAE). If we standarized MAE in terms of log mean is about 1.14%. MAE is 0.1278416, RSME accounts for 0.0297677 and 18.664869.