Homework Assignment #1

Quilvio Hernandez

4/2/2020

worked with: Aman Singh

Question 1

a. Use the substr and as.numeric function in R to generate new variables representing the year and month of the closing date.

```
Davis2018$CloseYear<-substring(Davis2018$ClosingDate,1,4)
Davis2018$CloseYear<-as.numeric(Davis2018$CloseYear)
Davis2018$CloseMonth<-substring(Davis2018$ClosingDate,6,7)
Davis2018$CloseMonth<-as.numeric(Davis2018$CloseMonth)

summary(Davis2018$CloseYear)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2018    2018    2018    2018    2018    2019
```

```
summary(Davis2018$CloseMonth)
```

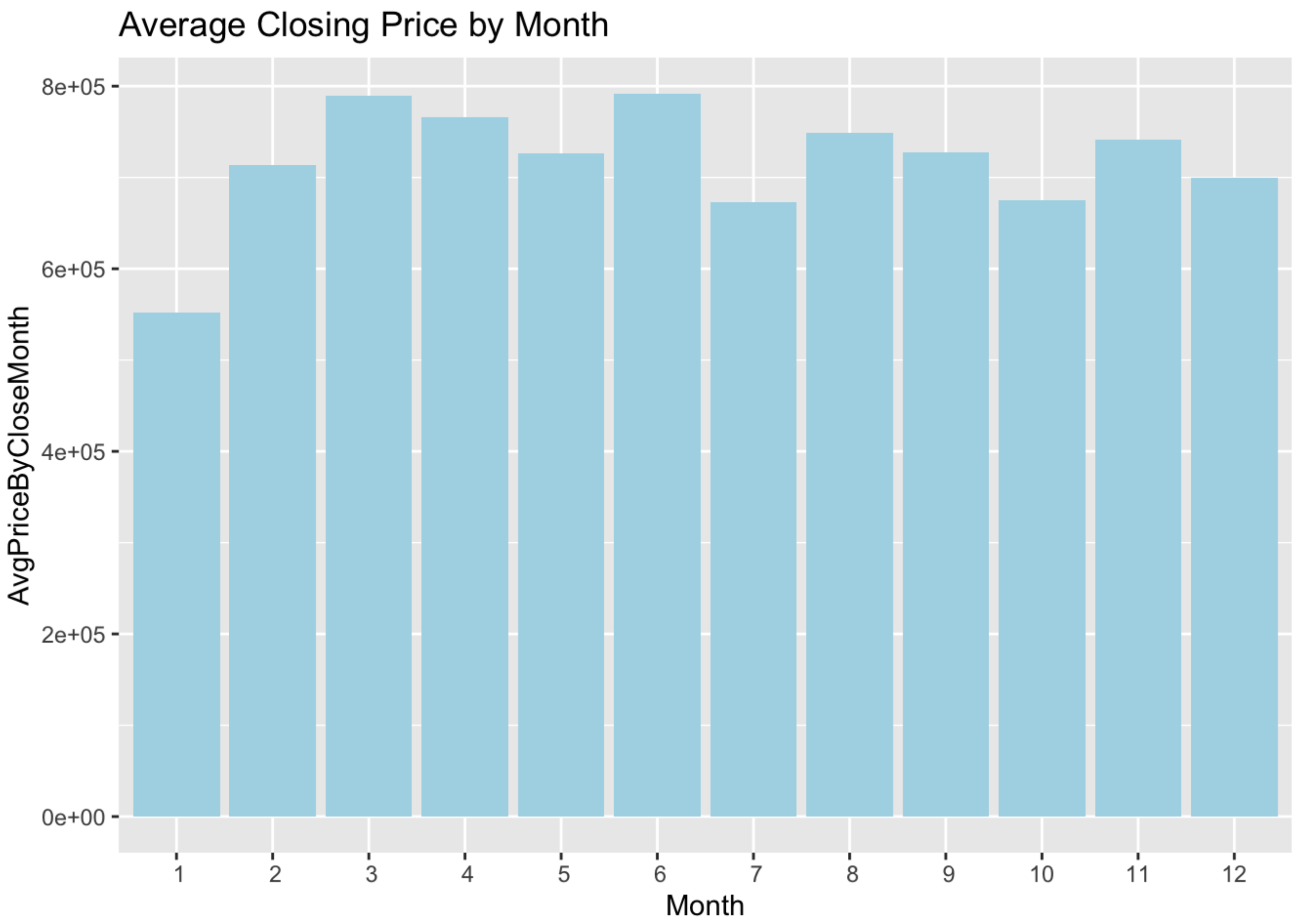
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   3.000   6.000   5.933   8.000  12.000
```

b. Restrict the sample to sales of single-family houses with close dates in 2018.

```
Davis2018_SF<- Davis2018[which(Davis2018$SingleFamily==1 & Davis2018$CloseYear==2018),]
```

c. Draw a bar plot to summarize the average sale price of houses with different characteristics of your choice (e.g., bedrooms, bathrooms, closing month, etc.) using the subsample created in part b.

```
AvgPriceByCloseMonth<-tapply(Davis2018_SF$SalePrice, Davis2018_SF$CloseMonth, mean)
library(ggplot2)
ggplot(data.frame(AvgPriceByCloseMonth),aes(seq_along(AvgPriceByCloseMonth),AvgPriceByCl
oseMonth)) +
  geom_bar(stat="identity", fill = "lightblue") +
  labs(x = "Month", title = "Average Closing Price by Month") +
  scale_x_discrete(limits=seq_along(AvgPriceByCloseMonth))
```



Average Closing Price by Month

d. Run a regression of sale price on month of closing and test the overall significance of the regression with 5% significance level.

```
summary(lm(SalePrice~factor(CloseMonth), data=Davis2018_SF))
```

```
##
## Call:
## lm(formula = SalePrice ~ factor(CloseMonth), data = Davis2018_SF)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -355496 -125903  -40930   91294  658562
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)            552125      72091   7.659 7.88e-13 ***
## factor(CloseMonth)2    161153      99080   1.626  0.10542
## factor(CloseMonth)3    237298      88293   2.688  0.00780 **
## factor(CloseMonth)4    213371      81136   2.630  0.00921 **
## factor(CloseMonth)5    173946      83695   2.078  0.03895 *
## factor(CloseMonth)6    239723      81743   2.933  0.00375 **
## factor(CloseMonth)7    120296      85938   1.400  0.16312
## factor(CloseMonth)8    197208      82826   2.381  0.01820 *
## factor(CloseMonth)9    174901      86642   2.019  0.04486 *
## factor(CloseMonth)10   122969      88293   1.393  0.16525
## factor(CloseMonth)11   189312     101952   1.857  0.06480 .
## factor(CloseMonth)12   147183      93069   1.581  0.11536
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 203900 on 200 degrees of freedom
## Multiple R-squared:  0.06627,    Adjusted R-squared:  0.01491
## F-statistic:  1.29 on 11 and 200 DF,  p-value: 0.232
```

With a p-value of 0.232, which is greater than 0.05, we fail to reject the null hypothesis and cannot conclude CloseMonth is significant at the 5% significance level.

e. How would you obtain heteroskedastic robust standard errors in the above regression if you think the homoskedasticity assumption is violated?

```
library(sandwich)
library(lmtest)
```

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

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

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

```
coeftest(lm(SalePrice~factor(CloseMonth), data=Davis2018_SF), vcov=sandwich)
```

```
##
## t test of coefficients:
##
##                        Estimate Std. Error t value  Pr(>|t|)
## (Intercept)              552125      57549  9.5940 < 2.2e-16 ***
## factor(CloseMonth)2      161153      68344  2.3580  0.019341 *
## factor(CloseMonth)3      237298      75482  3.1438  0.001922 **
## factor(CloseMonth)4      213371      69040  3.0905  0.002283 **
## factor(CloseMonth)5      173946      64663  2.6900  0.007748 **
## factor(CloseMonth)6      239723      71877  3.3352  0.001016 **
## factor(CloseMonth)7      120296      68825  1.7478  0.082026 .
## factor(CloseMonth)8      197208      72638  2.7149  0.007209 **
## factor(CloseMonth)9      174901      66094  2.6462  0.008787 **
## factor(CloseMonth)10     122969      79394  1.5488  0.122999
## factor(CloseMonth)11     189312     109597  1.7274  0.085648 .
## factor(CloseMonth)12     147183      92403  1.5928  0.112775
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

f. Run a regression of sale price on list price and days on market. How do you interpret the slope coefficients of this regression? Do you think the zero conditional mean condition is satisfied here?

```
results_List_Days<- lm(SalePrice~ListPrice+DaysOnMarket, data=Davis2018_SF)
summary(results_List_Days)
```

```
##
## Call:
## lm(formula = SalePrice ~ ListPrice + DaysOnMarket, data = Davis2018_SF)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -238348  -10025   -3568   10898   81732
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.185e+04  7.513e+03   2.908  0.00403 **
## ListPrice     9.839e-01  1.024e-02  96.081  < 2e-16 ***
## DaysOnMarket -3.853e+02  7.638e+01  -5.044 9.87e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 30060 on 209 degrees of freedom
## Multiple R-squared:  0.9788, Adjusted R-squared:  0.9786
## F-statistic:  4822 on 2 and 209 DF,  p-value: < 2.2e-16
```

The slope coefficients represent the partial effects of ListPrice and DaysOnMarket. That is to say for each additional unit increase in ListPrice we can expect, on average, an increase of $0.984 in SalePrice. Similarly, for each additional day the house is on the market we can expect, on average, a decrease of -$385.3 in SalePrice. I think the zero conditional mean condition is not satisfied here since DaysOnMarket and ListPrice are not independent of SalePrice. I think it's reasonable to believe SalePrice could have an (causal) effect on DaysOnMarket and ListPrice.

g. Add house characteristics to the above regression model and test the joint significance of all newly added house characteristics variables.

```
results_all<-lm(SalePrice~ListPrice+DaysOnMarket+Bedroom+Size+LotSize+FullBath+HalfBath+
HasPool+YearBuilt+Stories, data=Davis2018_SF)
summary(results_all)$adj.r.squared
```

```
## [1] 0.97912
```

```
summary(results_List_Days)$adj.r.squared
```

```
## [1] 0.9785871
```

```
qf(0.95, 3, 200)
```

```
## [1] 2.649752
```

```
library(car)
```

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

```
linearHypothesis(results_all,c("Bedroom=0","Size=0","LotSize=0","FullBath=0","HalfBath=
0","HasPool","YearBuilt","Stories"))
```

```
## Linear hypothesis test
##
## Hypothesis:
## Bedroom = 0
## Size = 0
## LotSize = 0
## FullBath = 0
## HalfBath = 0
## HasPool = 0
## YearBuilt = 0
## Stories = 0
##
## Model 1: restricted model
## Model 2: SalePrice ~ ListPrice + DaysOnMarket + Bedroom + Size + LotSize +
##     FullBath + HalfBath + HasPool + YearBuilt + Stories
##
##   Res.Df        RSS Df  Sum of Sq      F Pr(>F)
## 1    209 1.8889e+11
## 2    201 1.7713e+11  8 1.1751e+10 1.6667 0.1084
```

```
anova(results_all, results_List_Days)
```

```
## Analysis of Variance Table
##
## Model 1: SalePrice ~ ListPrice + DaysOnMarket + Bedroom + Size + LotSize +
##     FullBath + HalfBath + HasPool + YearBuilt + Stories
## Model 2: SalePrice ~ ListPrice + DaysOnMarket
##   Res.Df        RSS Df   Sum of Sq      F Pr(>F)
## 1    201 1.7713e+11
## 2    209 1.8889e+11 -8 -1.1751e+10 1.6667 0.1084
```

We fail to reject the null hypothesis that the house characteristics are joint significant.

h. Review your ECN 102 (or STA 108, ECN 140, etc…) notes on regressions with quadratic terms. Now, add a quadratic term of DaysOnMarket to the regression in f. For houses with the same list prices, what is the predicted difference in sale price if a house stays on market a week longer than the other?

```
quad_lm <- summary(lm(SalePrice~ListPrice+DaysOnMarket+I(DaysOnMarket^2), data=Davis2018
_SF))
quad_lm$coefficients
```

```
##                        Estimate   Std. Error   t value      Pr(>|t|)
## (Intercept)       23301.9935278 7357.9732913  3.166904  1.772318e-03
## ListPrice             0.9898581    0.0101765 97.269027 2.211749e-175
## DaysOnMarket       -932.6247971  183.3521494 -5.086522  8.126133e-07
## I(DaysOnMarket^2)     4.4528643    1.3623170  3.268596  1.264610e-03
```

```
week_diff <- function(ListPrice, DaysOnMarket = 0){
  dayzero = 23301.99 + 0.989*ListPrice - 932.62*(DaysOnMarket) + 4.45*(DaysOnMarket)^2
  dayweek = 23301.99 + 0.989*ListPrice - 932.62*(DaysOnMarket+7) + 4.45*(DaysOnMarket+7)
^2
  dayzero-dayweek
}
week_diff(100)
```

```
## [1] 6310.29
```

```
week_diff(1000000)
```

```
## [1] 6310.29
```

```
week_diff(1000000, 7)
```

```
## [1] 5874.19
```

```
week_diff(100, 7)
```

```
## [1] 5874.19
```

Fro houses with the same list prices, the predicted difference in sale price if a house stays on market a week longer than the other is $-932.62 * (DaysOnMarket + 7) + 4.45 * (DaysOnMarket + 7)^2$. The first two examples with the 6310.29 output show the fact that two pairs of houses with vastly different ListPrice (two with 100 vs. two with 1000000) display the same price difference after a week (6310.29$). The last two examples, with outputs 5874.19, show the nonlinearity effect of DaysOnMarket by starting with the initial condition that both pairs of houses have been on the market for a week, we ask what is the price change between the first and second week. One house in each pair is sold at day 7 while the other is sold at day 14 and the price difference we find is 5874.19. Again, we see that the price difference is the same between the pairs of houses despite their vastly different ListPrice.

Question 2

Use the RENTAL.dta dataset. This dataset comes from the Wooldridge textbook. It includes rental prices and other variables of 64 college towns for the years of 1980 and 1990.

```
rm(list = ls())
rental <- read_dta("/Users/quilviohernandez/Desktop/Spring2020/ECN 190/Data/RENTAL.DTA")
```

a. Review your ECN 102 (or STA 108, ECN 140, etc…) notes on regressions with log transformed variables. Regress log of *rent (lrent)* on log of *pop (lpop)*, log of *avginc (lavginc)*, and *pctstu* using only 1990 data. Interpret the slope coefficient of *lavginc* as well as pctstu. Do you think the zero conditional mean assumption is satisfied here?

```
summary(lm(lrent~lpop+ lavginc+ pctstu, data=rental, subset=(year==90)))$coefficients
```

```
##                Estimate  Std. Error  t value     Pr(>|t|)
## (Intercept) 0.042780259 0.843875331 0.050695 9.597370e-01
## lpop        0.065867849 0.038825998 1.696488 9.497630e-02
## lavginc     0.507015040 0.080835591 6.272176 4.294120e-08
## pctstu      0.005629696 0.001742066 3.231619 2.000534e-03
```

For each additional unit of *lavginc* we expect to see, on average, a 0.507 increase in *lrent*. Since log transformed variables can be interpreted as % change, we can say that if *lpop* were to increase by one log unit (or 100%) then we would expect *lrent* to increase by .507 (50.7% increase in rent), on average. If *pctstu* were to increase by one percent, *lrent* would increase by .006 (.6% increase in rent), on average. I think the zero conditional mean assumption is NOT satisfied because while all the independent variables appear to affect *rent*, I believe there are other causal macroeconic variables not present in the data such as unemployment, inflation, etc.

b. The variable *clrent* only has non-missing values in 1990. Verify those values are equal to the change in *lrent* in each city between year 1980 and year 1990. Recall that changes in log transformed variables could be interpreted as % changes in the original variable. Notice that *clrent* is equal to .5516071 for city 1. How do you interpret this number?

```
year80 <- subset(rental, year == 80)
year90 <- subset(rental, year == 90)
dim(year80)
```

```
## [1] 64 23
```

```
dim(year90)
```

```
## [1] 64 23
```

```
all((year90$lrent - year80$lrent) == year90$clrent)
```

```
## [1] TRUE
```

Since log trasformed variables could be interpreted as % change, we can subtract *lrent* from when year equals 90 and when the year is equal to 80 for each city to verify that *clrent* is equal to the change in *lrent* for each city. We can interpret *clrent*=0.5516071 in city 1 as a 55.16% increase in the average rent price between 1980 and 1990 in city 1.

c. Finally, we regress change in *lrent (clrent)* on change in *lpop (clpop)*, change in *lavginc (clavginc)*, and change in *pctstu (cpctstu)* between year 1980 and year 1990. How do you interpret the intercept here? Explain what the zero conditional mean assumption is requiring in this regression.

```
summary(lm(clrent~clpop+ clavginc+ cpctstu, data=rental))$coefficients
```

```
##                Estimate  Std. Error    t value     Pr(>|t|)
## (Intercept) 0.38552140 0.036824477 10.4691617 3.661078e-15
## clpop       0.07224555 0.088342598  0.8177884 4.167138e-01
## clavginc    0.30996054 0.066477145  4.6626633 1.788442e-05
## cpctstu     0.01120333 0.004131941  2.7113956 8.726352e-03
```

The intercept would be the change in *lrent (clrent)* (.386) should there be no change in *lpop (clpop = 0)*, *lavginc (clavginc = 0)*, *pctstu (cpctstu = 0)*. Since log trasformed variables could be interpreted as % change, this can be interpreted as a 38.6% increase in rent despite no changes in population, average income, or student population. The zero conditional mean assumption in this regression requires that the mean error terms conditioned on *clpop*, *clavginc*, and *cptstu* is 0. I believe this condition is not satisfied because there are other underlying macroeconomic conditions and welfare policies that affect *lrent*.