ECN190 Homework 1 Computer Problems

Kevin Chen (914861432) John Mayhew (914807483)

4/3/2020

1. Use Davis2018.dta.

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

```
##    ClosingDate ClosingYear ClosingMonth
## 1   2018-11-05        2018           11
## 2   2018-10-31        2018           10
## 3   2018-12-27        2018           12
## 4   2018-10-31        2018           10
## 5   2018-09-07        2018            9
## 6   2018-01-10        2018            1
## 7   2018-09-21        2018            9
## 8   2018-06-13        2018            6
## 9   2018-09-21        2018            9
## 10  2018-10-30        2018           10
```

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

```
##    ClosingYear SingleFamily
## 1         2018            1
## 2         2018            1
## 3         2018            1
## 4         2018            1
## 5         2018            1
## 6         2018            1
## 7         2018            1
## 8         2018            1
## 9         2018            1
## 10        2018            1
```

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.

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


```
##
## Call:
## lm(formula = SalePrice ~ ClosingMonth, data = Davis2018)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -340025 -135369  -25100  105842  676131
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    741640      34595  21.438   <2e-16 ***
## ClosingMonth    -1616       4907  -0.329    0.742
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 205900 on 210 degrees of freedom
## Multiple R-squared:  0.0005159,  Adjusted R-squared:  -0.004244
## F-statistic: 0.1084 on 1 and 210 DF,  p-value: 0.7423
```

It appears that regressing Closing Month on Sale Price is not effective; the p-value is 0.7423, far from significant at the 5% level.

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

We would carry out the t-test to obtain heteroskadasticity robust standard errors and their t-values.

```
##
## t test of coefficients:
##
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  741640.2    35131.0  21.111   <2e-16 ***
## ClosingMonth  -1615.6     5227.7  -0.309   0.7576
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

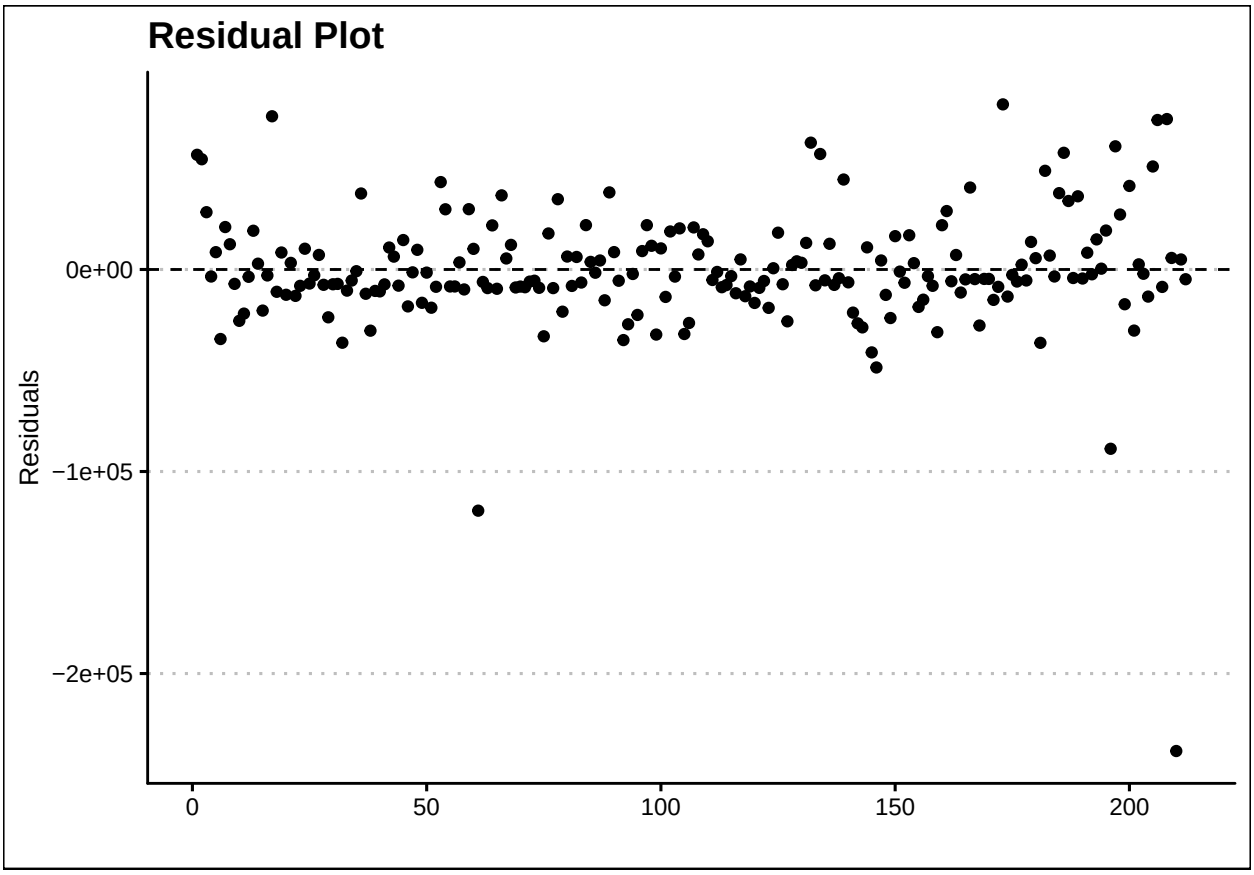
f. Run a regression of sale price on list price and days on market. How do you interpret the slope coefficients of this regression?

```
##
## Call:
## lm(formula = SalePrice ~ ListPrice + DaysOnMarket, data = Davis2018)
##
## Coefficients:
##  (Intercept)     ListPrice  DaysOnMarket
##   21846.3185        0.9839     -385.2674
```

The slope coefficients represent the increase or decrease in the dependent variable based on values of the independent variables; in this case, the coefficient for ListPrice (0.9839) is the increase in SalePrice with every 1-unit increase in ListPrice, and the coefficient for DaysOnMarket (-385.2674) is the decrease in SalePrice associated with every 1-unit increase in DaysOnMarket.

In other words, the ListPrice is slightly lower than SalePrice at every level, but mirrors its characteristics, while for every day on the market, a house loses around 385 dollars in value.

Do you think the zero conditional mean condition is satisfied here?



From plotting the residuals, we can see on average that the zero conditonal mean condition is satisfied because the residuals hover between -50000 and 50000, and the expected value is constant at all levels. There are a few outliers but on average, the residuals are 0 if they were to be summed up, allowing us to conclude a zero conditional mean.

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

```
##
## Call:
## lm(formula = SalePrice ~ ListPrice + DaysOnMarket + Bedroom +
##     FullBath + Stories, data = Davis2018)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -197000  -13049   -2131   11838   81446
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.036e+04  1.070e+04   0.968   0.3343
## ListPrice     1.006e+00  1.578e-02  63.744  < 2e-16 ***
## DaysOnMarket -3.493e+02  7.766e+01  -4.497 1.17e-05 ***
## Bedroom3      4.209e+03  8.072e+03   0.522   0.6026
## Bedroom4     -8.933e+03  9.595e+03  -0.931   0.3530
## Bedroom5     -2.274e+03  1.186e+04  -0.192   0.8481
## Bedroom6     -9.588e+03  2.310e+04  -0.415   0.6785
## Bedroom7      3.585e+04  3.359e+04   1.067   0.2872
## FullBath2     2.428e+03  7.404e+03   0.328   0.7433
## FullBath3     8.884e+03  1.001e+04   0.888   0.3757
## FullBath4    -4.047e+04  1.661e+04  -2.437   0.0157 *
## FullBath5     2.536e+04  2.570e+04   0.986   0.3251
## Stories      -5.002e+03  5.859e+03  -0.854   0.3942
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 29190 on 199 degrees of freedom
## Multiple R-squared:  0.981,  Adjusted R-squared:  0.9798
## F-statistic: 854.5 on 12 and 199 DF,  p-value: < 2.2e-16
```

From our original regression, we added 3 house characteristics, number of bedrooms, number of fullbaths, and stories. With p value $<$ 2.2e-16, which is close to 0, we can conclude there is joint significance of the newly added housing variables.

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?

```
##
## Call:
## lm(formula = SalePrice ~ ListPrice + DaysOnMarket + I(DaysOnMarket^2),
##     data = Davis2018)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -235156  -10821   -4620   12669   81425
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)        2.330e+04  7.358e+03   3.167  0.00177 **
## ListPrice          9.899e-01  1.018e-02  97.269  < 2e-16 ***
## DaysOnMarket      -9.326e+02  1.834e+02  -5.087 8.13e-07 ***
## I(DaysOnMarket^2)  4.453e+00  1.362e+00   3.269  0.00126 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 29390 on 208 degrees of freedom
## Multiple R-squared:  0.9798, Adjusted R-squared:  0.9795
## F-statistic:  3367 on 3 and 208 DF,  p-value: < 2.2e-16
```

Quadratic Model: 23301.9935 + ListPrice(0.9889) + DaysOnMarket(-932.6248) + DaysOnMarketˆ2(4.4529)

If two houses have identical List Prices, and one stays on the market 1 week longer:

Let C = 23301.9935 + ListPrice(0.9889)

House 1 = C + 0

House 2 = C + 7*(-932.6248) + 49*(4.4529) = C - 6310.182

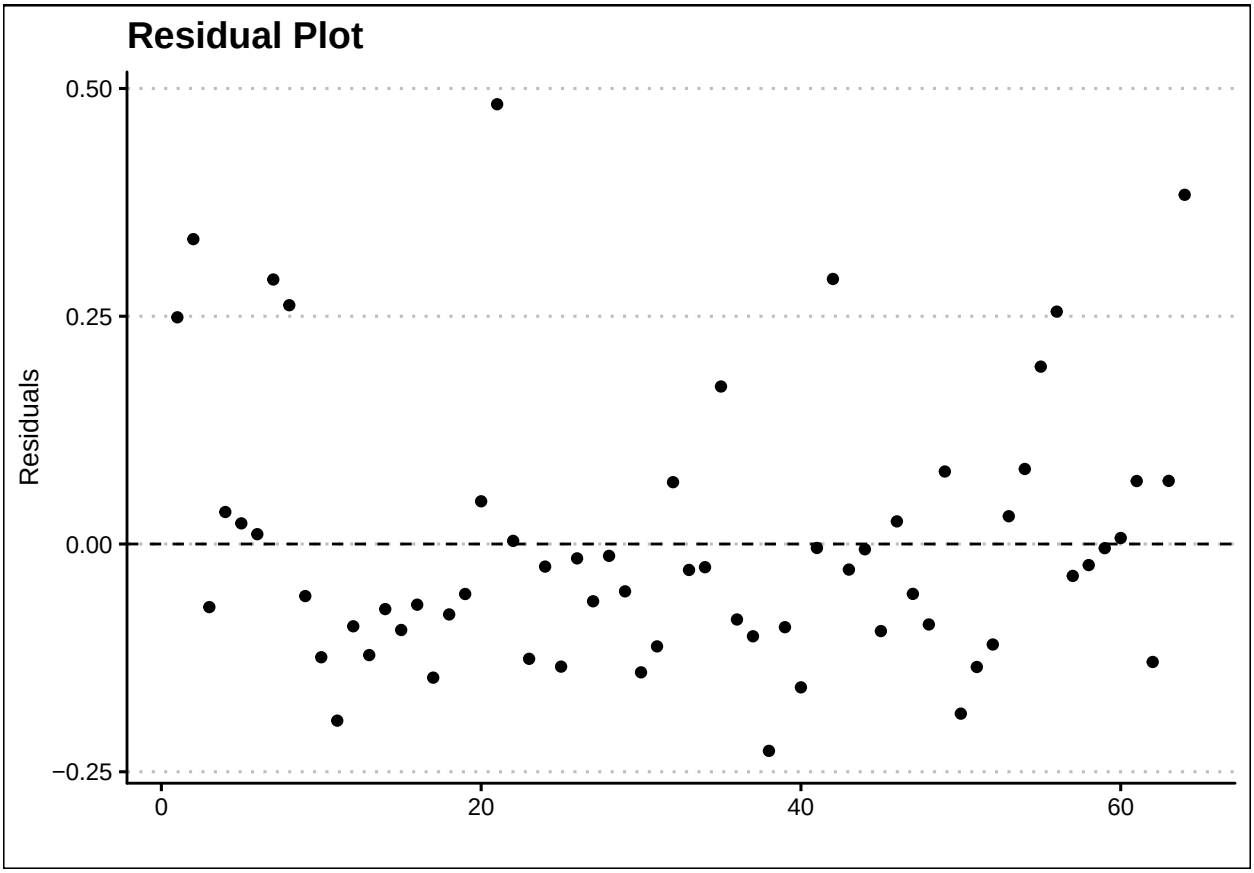
"The house that stays on the market 7 days longer loses $6,310.18 in value."

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.

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?

```
##
## Call:
## lm(formula = lrent ~ lpop + lavginc + pctstu, data = rental1990)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.22706 -0.09469 -0.02827  0.03806  0.48271
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.042780   0.843875   0.051    0.960
## lpop        0.065868   0.038826   1.696    0.095 .
## lavginc     0.507015   0.080836   6.272 4.29e-08 ***
## pctstu      0.005630   0.001742   3.232    0.002 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1512 on 60 degrees of freedom
## Multiple R-squared:  0.4579, Adjusted R-squared:  0.4308
## F-statistic: 16.89 on 3 and 60 DF,  p-value: 4.541e-08
```

Since this model is a log regression, we can interpret the coefficients as percentages/elasticities. If we change lavginc (log of average income) by 1%, we would expect rent to change by 0.5%. However, for pctstu (percentage of student), we did not take the log of it since it is in percentages already. So if we change pctstu by 1 unit (% in this case), we would expect rent to change by 0.563%.



For this model, it does not seem that the zero conditional mean assumption is satisfied here as the majority of the residuals are negative which means on average, the mean of the residuals is not zero.

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?

For city 1, the clrent is equal to 0.5516071. This means that in city 1, the rent in 1990 was 55.16% higher than it was in 1980, or there was a 55.16% change in rent from 1980 to 1990.

```
##    city year    clrent    lrent rent clrent.calc
## 2     1   90 0.5516071 5.834811  342   0.5516071
## 4     2   90 0.4289236 6.206576  496   0.4289236
## 6     3   90 0.4855080 5.860786  351   0.4855080
## 8     4   90 0.7894783 6.376727  588   0.7894783
## 10    5   90 0.6664791 6.829794  925   0.6664791
## 12    6   90 0.8253188 6.445720  630   0.8253188
## 14    7   90 0.5453234 6.255750  521   0.5453234
## 16    8   90 0.4959292 6.045005  422   0.4959292
## 18    9   90 0.8087320 6.342122  568   0.8087320
## 20   10   90 0.4590969 5.948035  383   0.4590969
```

```
all(rentaldata2b.omit$clrent == rentaldata2b.omit$clrent.calc)
```

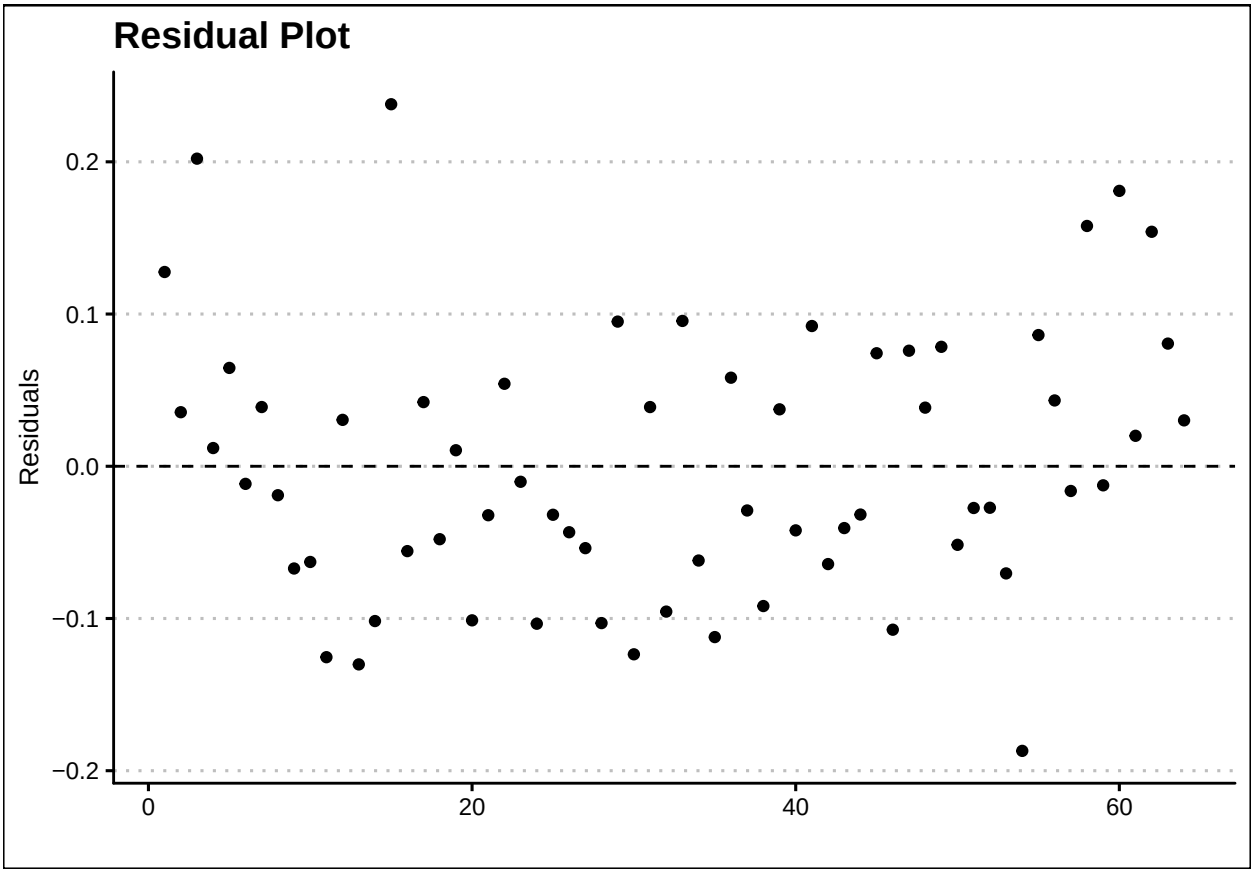
```
## [1] TRUE
```

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.

```
##
## Call:
## lm(formula = clrent ~ clpop + clavginc + cpctstu, data = rentaldata.omit)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.18697 -0.06216 -0.01438  0.05518  0.23783
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.385521   0.036824  10.469 3.66e-15 ***
## clpop       0.072246   0.088343   0.818  0.41671
## clavginc    0.309961   0.066477   4.663 1.79e-05 ***
## cpctstu     0.011203   0.004132   2.711  0.00873 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.09013 on 60 degrees of freedom
## Multiple R-squared:  0.3223, Adjusted R-squared:  0.2884
## F-statistic:  9.51 on 3 and 60 DF,  p-value: 3.136e-05
```

The intercept (0.385521) is the percent change in rent that would occur without any change in population, average income, or percentage of students; even if nothing else in the model changes, the rent would still increase by around 38.5%.

In the context of this regression, the zero conditional mean assumption requires that the expected differnence between the actual percent change in rent and the predicted percent change in rent based on our variables has a mean of 0, meaning that the residual plot of the errors should be randomly distributed about the y-intercept.



It appears from the plot that the zero-conditional mean is satisfied. The majority of the residuals hover between -0.2 and 0.2, so on average, the sum of the residuals equate to zero. This leads us to the conclusion that the zero-conditional mean is satisfied from looking at the residual plot.

Appendix

```
library(readstata13)
library(dplyr)
library(ggplot2)
library(ggpubr)
library(lmtest)
library(sandwich)
library(ggthemes)
library(knitr)
Davis2018 = read.dta13("Davis2018.dta")
Davis2018$ClosingYear <- as.numeric(substr(Davis2018$ClosingDate,1,4))
Davis2018$ClosingMonth <- as.numeric(substr(Davis2018$ClosingDate,6,7))
p1a.output =
  Davis2018 %>% select(ClosingDate, ClosingYear, ClosingMonth)
head(p1a.output, 10)

Davis2018 = filter(Davis2018, ClosingYear == 2018 & SingleFamily == 1)
p2a.output =
```

```
  Davis2018 %>% select(ClosingYear, SingleFamily)
head(p2a.output, 10)

Davis2018$Bedroom = as.factor(Davis2018$Bedroom)
tab = Davis2018 %>% group_by(Bedroom) %>% summarise(mean(SalePrice))
tab = data.frame(tab)

names(tab) = c("NumberBedrooms", "AvgSale")
pl = ggplot(data = tab, aes(x = NumberBedrooms, y = AvgSale)) +
  geom_bar(stat = "identity", width = I(1/3), color = "black", fill = "#F7CAC9") + theme_bw() +
  scale_x_discrete("Number of Bedrooms") +
  scale_y_continuous("Average Sale Price",
                     breaks = c(0, 250000, 500000, 750000, 1000000, 1250000, 1500000),
                     limits = c(0, 1570000)) +
  ggtitle("Avg. Sale Price by Bedrooms")

Davis2018$FullBath = as.factor(Davis2018$FullBath)
tab = Davis2018 %>% group_by(FullBath) %>% summarise(mean(SalePrice))
tab = data.frame(tab)

names(tab) = c("NumberFullBaths", "AvgSale")
pl2 = ggplot(data = tab, aes(x = NumberFullBaths, y = AvgSale)) +
  geom_bar(stat = "identity", width = I(1/5), color = "black", fill = "#F7CAC9") + theme_bw() +
  scale_x_discrete("Number of Full Bathrooms") +
  scale_y_continuous("Average Sale Price", breaks = c(0, 250000, 500000, 750000, 1000000,
                                                      1250000, 1500000), limits = c(0, 1570000)) +
  ggtitle("Avg. Sale Price by Bathrooms")

Davis2018$ClosingMonth = as.factor(Davis2018$ClosingMonth)
tab = Davis2018 %>% group_by(ClosingMonth) %>% summarise(mean(SalePrice))
tab = data.frame(tab)

names(tab) = c("NumberClosingMonths", "AvgSale")
pl3 = ggplot(data = tab, aes(x = NumberClosingMonths, y = AvgSale)) +
  geom_bar(stat = "identity", width = I(1/5), color = "black", fill = "#F7CAC9") + theme_bw() +
  scale_x_discrete("Month") +
  scale_y_continuous("Average Sale Price", breaks = c(0, 250000, 500000, 750000, 1000000),
                     limits = c(0, 1050000)) +
  ggtitle("Avg. Sale Price by Month")

Davis2018$Areacode = as.factor(Davis2018$Areacode)
tab = Davis2018 %>% group_by(Areacode) %>% summarise(mean(SalePrice))
tab = data.frame(tab)

names(tab) = c("NumberAreacodes", "AvgSale")
pl4 = ggplot(data = tab, aes(x = NumberAreacodes, y = AvgSale)) +
  geom_bar(stat = "identity", width = I(1/7), color = "black", fill = "#F7CAC9") + theme_bw() +
  scale_x_discrete("Area Code") +
  scale_y_continuous("Average Sale Price", breaks = c(0, 250000, 500000, 750000, 1000000),
                     limits = c(0, 1050000)) +
  ggtitle("Avg. Sale Price by Area Code")

ggarrange(pl, pl2, pl3, pl4, ncol = 2, nrow = 2)
```

```r
Davis2018$ClosingMonth = as.double(Davis2018$ClosingMonth)
lm.model = lm(SalePrice ~ ClosingMonth, data = Davis2018)
ggplot(data = Davis2018, aes(x = ClosingMonth, y = SalePrice)) + geom_point() +
  theme_bw() + scale_x_continuous("Closing Month", breaks = seq(1, 12, 1), limits = c(0.75, 12.25)) +
  scale_y_continuous("Sale Price", breaks = seq(0, 1500000, by = 150000), limits = c(0, 1550000)) +
  geom_smooth(method = "lm", se = F, formula = "y ~ x") + ggtitle("Closing Month versus Sale Price")

summary(lm.model)
coeftest(lm.model, vcov = sandwich)
Davis2018$DaysOnMarket = as.double(Davis2018$DaysOnMarket)
Davis2018$ListPrice = as.double(Davis2018$ListPrice)
lm.model2 = lm(SalePrice ~ ListPrice + DaysOnMarket, data = Davis2018)
print(lm.model2)
cat("\\newpage")
df1 = Davis2018 %>% select(ListPrice, DaysOnMarket, SalePrice)
df = cbind.data.frame(df1, lm.model2$residuals, lm.model2$fitted.values)
df =
  df %>% arrange(`lm.model2$fitted.values`) #%>% select()
df$Index = 1:nrow(df)
df =
  df %>% select(Index, `lm.model2$residuals`)
names(df) = c("Index", "Residuals")
ggplot(data = df, aes(x = Index, y = Residuals)) + geom_point() +
  theme_clean() + geom_hline(yintercept = 0, linetype = "dashed") +
  scale_x_continuous("") + ggtitle("Residual Plot")
cat("\\newpage")
lm.model3 = lm(SalePrice ~ ListPrice + DaysOnMarket + Bedroom + FullBath + Stories, data = Davis2018)
summary(lm.model3)
cat("\\newpage")
lm.model4 = lm(SalePrice ~ ListPrice + DaysOnMarket + I(DaysOnMarket^2), data = Davis2018)
summary(lm.model4)
cat("\\newpage")
rentaldata <- read.dta13("RENTAL.DTA")

rental1990 <- subset(rentaldata, year != 80)

model2 <- lm(lrent ~ lpop + lavginc + pctstu, data = rental1990)
summary(model2)
df2 <- rental1990 %>% select(lpop, lavginc, pctstu, lrent)
df2.1 <- cbind.data.frame(df2, model2$residuals, model2$fitted.values)
df2.1 <-
  df2.1 %>% arrange(`model2$fitted.values`) #%>% select()
df2.1$Index <- 1:nrow(df2.1)
df2.1 <-
  df2.1 %>% select(Index, `model2$residuals`)
names(df2.1) = c("Index", "Residuals")
ggplot(data = df2.1, aes(x = Index, y = Residuals)) + geom_point() +
  theme_clean() + geom_hline(yintercept = 0, linetype = "dashed") +
  scale_x_continuous("") + ggtitle("Residual Plot")
rentaldata2b =
  rentaldata %>% select(city, year, clrent, lrent, rent)

for (i in 1:nrow(rentaldata2b)){
```

```
  if (i %% 2 == 0){
    rentaldata2b$clrent.calc[i] = rentaldata2b$lrent[i] - rentaldata2b$lrent[i-1]

  } else{
    rentaldata2b$clrent.calc[i] = 0
  }
}
rentaldata2b.omit = na.omit(rentaldata2b)
print(head(rentaldata2b.omit, 10))

all(rentaldata2b.omit$clrent == rentaldata2b.omit$clrent.calc)
rentaldata.omit = na.omit(rentaldata)
rentaldata.omit =
  rentaldata.omit %>% select(clrent, clpop, clavginc, cpctstu)
lm.modelc = lm(clrent ~ clpop + clavginc + cpctstu, data = rentaldata.omit)

summary(lm.modelc)
df3 = rentaldata.omit
df3.1 <- cbind.data.frame(df3, lm.modelc$residuals, lm.modelc$fitted.values)
df3.1 <-
  df3.1 %>% arrange(`lm.modelc$fitted.values`)
df3.1$Index <- 1:nrow(df3.1)
df3.1 <-
  df3.1 %>% select(Index, `lm.modelc$residuals`)
names(df3.1) = c("Index", "Residuals")
ggplot(data = df3.1, aes(x = Index, y = Residuals)) + geom_point() +
  theme_clean() + geom_hline(yintercept = 0, linetype = "dashed") +
  scale_x_continuous("") + ggtitle("Residual Plot")
```