Regression\_Project\_Final

Aaron Goletz

December 6, 2017

# 1) An examination of the distributions of the variables

# comm\_prop <- read.csv("C:/Users/aaron.goletz/Desktop/My Docs/Lipscomb Documents/Stats Analysis and Decision Modeling/MSDS 5043/comm\_prop.csv")  
comm\_prop <- read.csv("comm\_prop.csv", header=TRUE)  
comm\_prop$W2MiDT <- factor(comm\_prop$W2MiDT, labels = c("No", "Yes"))  
  
kable(head(comm\_prop), format = "html", align='c', caption = "Example of Data") #Head function will return the FIRST 6 rows of the dataset

Example of Data

RentRate

Age

OperExp

VacRate

SqFt

Taxes

W2MiDT

13.5

1

5.02

0.14

123000

2.21

No

12.0

14

8.19

0.27

104079

1.87

No

10.5

16

3.00

0.00

39998

0.72

No

15.0

4

10.70

0.05

57112

1.03

No

14.0

11

8.97

0.07

60000

1.08

No

10.5

15

9.45

0.24

101385

1.82

No

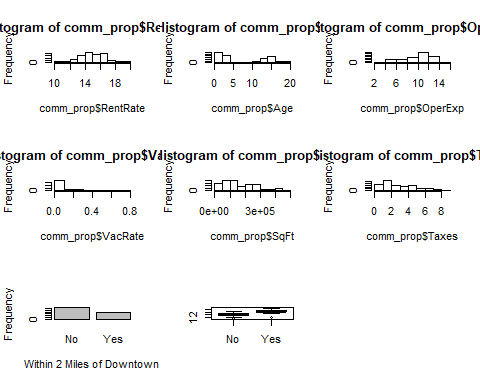
# tail(comm\_prop) #Tail function will return the LAST 6 rows of the dataset  
summary(comm\_prop) #Summary function returns a compact analysis of the dataset

## RentRate Age OperExp VacRate   
## Min. :10.50 Min. : 0.00 Min. : 3.000 Min. :0.0000   
## 1st Qu.:14.00 1st Qu.: 2.00 1st Qu.: 7.997 1st Qu.:0.0000   
## Median :15.00 Median : 4.00 Median :10.370 Median :0.0400   
## Mean :15.13 Mean : 8.01 Mean : 9.658 Mean :0.0801   
## 3rd Qu.:16.50 3rd Qu.:15.00 3rd Qu.:11.660 3rd Qu.:0.0925   
## Max. :19.25 Max. :20.00 Max. :14.620 Max. :0.7300   
## SqFt Taxes W2MiDT   
## Min. : 27000 Min. :0.490 No :64   
## 1st Qu.: 73141 1st Qu.:1.315 Yes:36   
## Median :129614 Median :2.330   
## Mean :166101 Mean :2.990   
## 3rd Qu.:244547 3rd Qu.:4.402   
## Max. :484290 Max. :8.720

Observations: Head, Tail and Summary metrics gives sense of the data. Just by looking at the brief summary, we can notice few variables like SqFt, Taxes has a higher mean than its median. This indicates the variables are probably right skewed. We will explore further with histogram plots.

Variables are : RentRate - Continuous Variable Age - Continuous Variable OperExp - Continuous Variable VacRate - Continuous Variable SqFt - Continuous Variable Taxes - Continuous Variable W2MiDT - Categorical Variable

### Histograms for distribution visualization



Observations: Histograms help us better understand the distribution of continuous variables. Above, we have plotted all continuous variables into histograms and bar plot and boxplot for Categorical Variable.

We noticed the following: Rental Rate - looks normally distributed Age - appears bimodal. It has an unusual distribution shape to it. It looks like there were buildings built recently and also around 15 years ago it was higher. OperExp - contains a slight amount of left/negative skewness. VacRate - contains strong Right/positive skewness SqFt - contain a slight amount of right skewness Taxes - contain a slight amount of right skewness W2MiDTNo - has a higher frequency.

# 2) An examination of the correlations for all continuous variables

**Correlation Matrix**

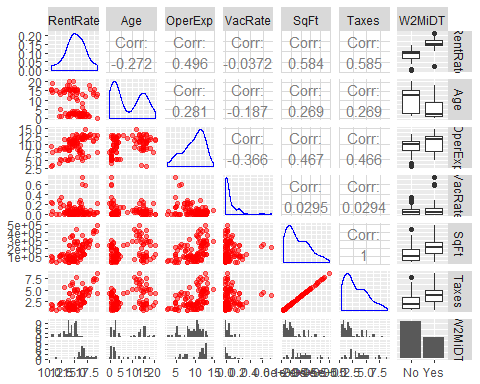
round(cor(comm\_prop[1:6]),3)

## RentRate Age OperExp VacRate SqFt Taxes  
## RentRate 1.000 -0.272 0.496 -0.037 0.584 0.585  
## Age -0.272 1.000 0.281 -0.187 0.269 0.269  
## OperExp 0.496 0.281 1.000 -0.366 0.467 0.466  
## VacRate -0.037 -0.187 -0.366 1.000 0.029 0.029  
## SqFt 0.584 0.269 0.467 0.029 1.000 1.000  
## Taxes 0.585 0.269 0.466 0.029 1.000 1.000

**Correlation Matrix with scatterplots and correlation coefficients**

ggpairs(comm\_prop[,1:7] ,   
 lower = list(continuous = wrap("points", color = "red", alpha = 0.5)),   
 diag = list(continuous = wrap("densityDiag", color = "blue", alpha = 0.5)))

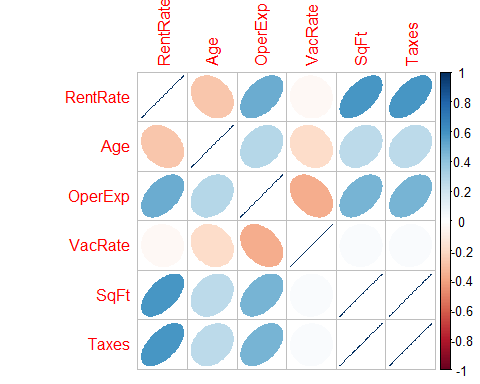
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.  
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.  
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.  
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.  
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.  
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



# 'ggpairs' function is derived from 'GGally' library and draws a martix of plots of the diven dataset

**Correlation Matrix with ellipse representation of scatterplots**

corrplot(cor(comm\_prop[,c(1:6)]), method="ellipse")



#'corrplot' function is derived from the 'corrplot' library and returns the correlations of multiple factors of the dataset in graphical table

Observations:  
SqFt and Taxes - have a perfect correlation of 1.00. RentRate and OperExp - have a moderate positive correlation. RentRate and SqFt - have a strong positive correlation.  
RentRate and Taxes - share the same correlation as of .584.

Other correlations to note are: VacRate and OperExp - a negative correlation of -.366 Age and OperExp - a very small correlation Age and RentRate - a very small negative correlation

# 3) The identification and evaluation of a suitable regression model for predicting rental rates

A manual iteration selection method will be used first.

Manual Iteration is a process of building the best-fit model by manually observing the data set and eliminating the factors without a significant p-value one by one. Then observing the new modelâs adjusted R-sqaure value. The process of elimination is done for every variable that has the highest p-value (i.e is LEAST significant).

For this, we use the âlmâ function, to fit Linear Models to the datasets and execute the âsummaryâ function on the model.

The results will later be compared to results from Stepwise and All Subsets Regression (ASR) selection methods in order to select the most appropriate linear model.

## a) FIT FULL MODEL

m <- lm(RentRate ~ ., data=comm\_prop)  
summary(m)

##   
## Call:  
## lm(formula = RentRate ~ ., data = comm\_prop)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.65450 -0.59760 -0.02213 0.65394 2.57988   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 12.2278960 0.4483644 27.272 < 2e-16 \*\*\*  
## Age -0.1040412 0.0185484 -5.609 2.08e-07 \*\*\*  
## OperExp 0.2255069 0.0497984 4.528 1.76e-05 \*\*\*  
## VacRate -0.7568956 0.9114326 -0.830 0.408   
## SqFt -0.0005652 0.0007194 -0.786 0.434   
## Taxes 31.7633824 39.9476685 0.795 0.429   
## W2MiDTYes 1.4434308 0.2586992 5.580 2.36e-07 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.9884 on 93 degrees of freedom  
## Multiple R-squared: 0.7463, Adjusted R-squared: 0.73   
## F-statistic: 45.61 on 6 and 93 DF, p-value: < 2.2e-16

Observations: Results show the model’s significant variables are Age, OperExp, and W2MiDT. The Adjusted R-squared value is good at 0.73, with an F-statistic of 45.61 and a high significant p-value of <2.2e-16. The model will now be trimmed to remove non-significant variables.

First, VacRate will be dropped due to non-significance in addition to it being an unknown variable in our final prediction. VacRate data is not as readily available as the others. Also.,the variable Taxes (0.429) will be dropped from the model, due to multicolinearity with SqFt (0.434).

## ITERATE MANUALLY

m.4 <- lm(RentRate ~ Age + OperExp + VacRate + SqFt + W2MiDT, data=comm\_prop) # drop Taxes  
summary(m.4)

##   
## Call:  
## lm(formula = RentRate ~ Age + OperExp + VacRate + SqFt + W2MiDT,   
## data = comm\_prop)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.73824 -0.60908 -0.05855 0.59708 2.49110   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.227e+01 4.442e-01 27.627 < 2e-16 \*\*\*  
## Age -1.050e-01 1.847e-02 -5.685 1.46e-07 \*\*\*  
## OperExp 2.175e-01 4.866e-02 4.469 2.19e-05 \*\*\*  
## VacRate -8.637e-01 8.997e-01 -0.960 0.34   
## SqFt 6.851e-06 1.133e-06 6.046 2.98e-08 \*\*\*  
## W2MiDTYes 1.468e+00 2.563e-01 5.729 1.21e-07 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.9864 on 94 degrees of freedom  
## Multiple R-squared: 0.7446, Adjusted R-squared: 0.731   
## F-statistic: 54.82 on 5 and 94 DF, p-value: < 2.2e-16

m.3 <- lm(RentRate ~ Age + OperExp + VacRate + SqFt, data=comm\_prop) # drop W2MiDT  
summary(m.3)

##   
## Call:  
## lm(formula = RentRate ~ Age + OperExp + VacRate + SqFt, data = comm\_prop)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.85766 -0.59144 -0.07254 0.60765 2.86660   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.215e+01 5.126e-01 23.701 < 2e-16 \*\*\*  
## Age -1.550e-01 1.881e-02 -8.244 9.19e-13 \*\*\*  
## OperExp 2.786e-01 5.485e-02 5.080 1.88e-06 \*\*\*  
## VacRate -1.698e-01 1.030e+00 -0.165 0.869   
## SqFt 9.288e-06 1.213e-06 7.655 1.60e-11 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.14 on 95 degrees of freedom  
## Multiple R-squared: 0.6555, Adjusted R-squared: 0.641   
## F-statistic: 45.18 on 4 and 95 DF, p-value: < 2.2e-16

m.2 <- lm(RentRate ~ Age + OperExp + VacRate + W2MiDT, data=comm\_prop) # drop SqFt  
summary(m.2)

##   
## Call:  
## lm(formula = RentRate ~ Age + OperExp + VacRate + W2MiDT, data = comm\_prop)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.9848 -0.6396 -0.0114 0.6751 2.9909   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 11.81786 0.51324 23.026 < 2e-16 \*\*\*  
## Age -0.06764 0.02041 -3.315 0.0013 \*\*   
## OperExp 0.32091 0.05341 6.009 3.43e-08 \*\*\*  
## VacRate 0.17672 1.03527 0.171 0.8648   
## W2MiDTYes 2.05008 0.27849 7.362 6.50e-11 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.156 on 95 degrees of freedom  
## Multiple R-squared: 0.6453, Adjusted R-squared: 0.6304   
## F-statistic: 43.21 on 4 and 95 DF, p-value: < 2.2e-16

m.1 <- lm(RentRate ~ Age + OperExp + SqFt + W2MiDT, data=comm\_prop) # drop VacRate  
summary(m.1)

##   
## Call:  
## lm(formula = RentRate ~ Age + OperExp + SqFt + W2MiDT, data = comm\_prop)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.90483 -0.58081 -0.03454 0.57347 2.54837   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.205e+01 3.789e-01 31.797 < 2e-16 \*\*\*  
## Age -1.040e-01 1.843e-02 -5.641 1.74e-07 \*\*\*  
## OperExp 2.373e-01 4.406e-02 5.385 5.22e-07 \*\*\*  
## SqFt 6.643e-06 1.112e-06 5.975 3.98e-08 \*\*\*  
## W2MiDTYes 1.435e+00 2.539e-01 5.653 1.65e-07 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.986 on 95 degrees of freedom  
## Multiple R-squared: 0.7421, Adjusted R-squared: 0.7313   
## F-statistic: 68.35 on 4 and 95 DF, p-value: < 2.2e-16

Observations: All variables are highly significant. Model m.1’s summary shows an increased Adjusted R-squared value of .7313, an increased F-Statistic of 68.35, and no notable change in the p-value.\*

## b) Stepwise Iterations

comm\_prop.full <- lm(RentRate ~ .\*., data = comm\_prop)  
stepAIC(comm\_prop.full, direction ="backward")

The stepAIC yields the following formula for an optimal linear model:

formula = RentRate ~ Age + OperExp + VacRate + Taxes + W2MiDT +   
Age:OperExp + Age:VacRate + OperExp:VacRate + OperExp:W2MiDT +   
Taxes:W2MiDT

m.step <- lm(formula = RentRate ~ Age + OperExp + VacRate + Taxes + W2MiDT +   
 Age:OperExp + Age:VacRate + OperExp:VacRate + OperExp:W2MiDT +   
 Taxes:W2MiDT, data = comm\_prop)  
summary(m.step)

##   
## Call:  
## lm(formula = RentRate ~ Age + OperExp + VacRate + Taxes + W2MiDT +   
## Age:OperExp + Age:VacRate + OperExp:VacRate + OperExp:W2MiDT +   
## Taxes:W2MiDT, data = comm\_prop)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.15120 -0.41853 0.02372 0.41115 2.07336   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 12.568031 0.556461 22.586 < 2e-16 \*\*\*  
## Age -0.189674 0.056318 -3.368 0.001121 \*\*   
## OperExp 0.253455 0.063470 3.993 0.000134 \*\*\*  
## VacRate -3.222338 2.527990 -1.275 0.205746   
## Taxes 0.076454 0.062327 1.227 0.223187   
## W2MiDTYes 2.117163 0.751416 2.818 0.005961 \*\*   
## Age:OperExp 0.011972 0.005558 2.154 0.033956 \*   
## Age:VacRate -0.542475 0.156238 -3.472 0.000799 \*\*\*  
## OperExp:VacRate 0.516211 0.381367 1.354 0.179298   
## OperExp:W2MiDTYes -0.253997 0.076312 -3.328 0.001272 \*\*   
## Taxes:W2MiDTYes 0.558834 0.093218 5.995 4.27e-08 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.7513 on 89 degrees of freedom  
## Multiple R-squared: 0.8597, Adjusted R-squared: 0.844   
## F-statistic: 54.55 on 10 and 89 DF, p-value: < 2.2e-16

Removing VacRate and associating interactions.

m.step <- lm(formula = RentRate ~ Age + OperExp + Taxes + W2MiDT +   
 Age:OperExp + OperExp:W2MiDT + Taxes:W2MiDT, data = comm\_prop)  
summary(m.step)

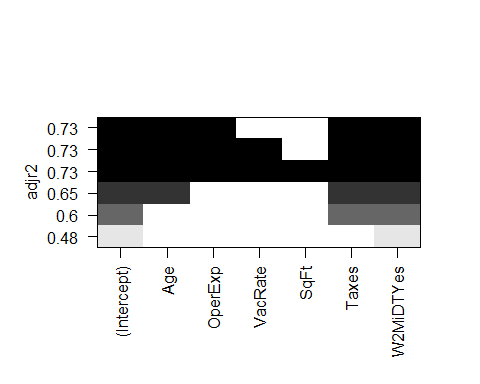
##   
## Call:  
## lm(formula = RentRate ~ Age + OperExp + Taxes + W2MiDT + Age:OperExp +   
## OperExp:W2MiDT + Taxes:W2MiDT, data = comm\_prop)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.02214 -0.45535 0.02514 0.50490 2.58140   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 12.472493 0.569956 21.883 < 2e-16 \*\*\*  
## Age -0.255072 0.056376 -4.524 1.81e-05 \*\*\*  
## OperExp 0.265460 0.065836 4.032 0.000114 \*\*\*  
## Taxes 0.072665 0.067086 1.083 0.281566   
## W2MiDTYes 1.925226 0.721427 2.669 0.009002 \*\*   
## Age:OperExp 0.015766 0.005845 2.697 0.008311 \*\*   
## OperExp:W2MiDTYes -0.226073 0.075188 -3.007 0.003404 \*\*   
## Taxes:W2MiDTYes 0.544866 0.099748 5.462 3.97e-07 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.8135 on 92 degrees of freedom  
## Multiple R-squared: 0.83, Adjusted R-squared: 0.8171   
## F-statistic: 64.18 on 7 and 92 DF, p-value: < 2.2e-16

The m.step model has an Adjusted R-squared value of 0.8171 and an F-Statistic of 64.18 with a highly significant p-value, <2.2e-16.

*The Taxes variable is not significant, but it will stay in the model as its interaction with W2MiDT is highly significant.*

## c) ASR Iteration

subsets <- regsubsets(RentRate ~ ., data=comm\_prop)  
plot(subsets, scale="adjr2")



Based on the subsets plot, our variable selection should be Age, OperExp, W2MiDT, Age:OperExp, Age:VacRate, Age:Taxes, OperExp:W2MiDT, and SqFt:W2MiDT.

No VacRate interaction will be used, as it is an unknown variable in our final prediction.

m.sub <- lm(RentRate ~ Age+OperExp+W2MiDT+Age:OperExp+Age:Taxes+OperExp:W2MiDT+SqFt:W2MiDT, data=comm\_prop)  
summary(m.sub)

##   
## Call:  
## lm(formula = RentRate ~ Age + OperExp + W2MiDT + Age:OperExp +   
## Age:Taxes + OperExp:W2MiDT + SqFt:W2MiDT, data = comm\_prop)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.95993 -0.47087 0.00436 0.43161 2.63782   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.243e+01 5.707e-01 21.773 < 2e-16 \*\*\*  
## Age -2.451e-01 5.702e-02 -4.298 4.31e-05 \*\*\*  
## OperExp 2.959e-01 7.125e-02 4.152 7.40e-05 \*\*\*  
## W2MiDTYes 2.297e+00 7.945e-01 2.891 0.00480 \*\*   
## Age:OperExp 1.193e-02 6.790e-03 1.756 0.08238 .   
## Age:Taxes 1.049e-02 9.469e-03 1.108 0.27072   
## OperExp:W2MiDTYes -2.594e-01 8.085e-02 -3.208 0.00184 \*\*   
## W2MiDTNo:SqFt -6.249e-07 2.120e-06 -0.295 0.76882   
## W2MiDTYes:SqFt 9.339e-06 2.186e-06 4.273 4.73e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.8125 on 91 degrees of freedom  
## Multiple R-squared: 0.8323, Adjusted R-squared: 0.8175   
## F-statistic: 56.45 on 8 and 91 DF, p-value: < 2.2e-16

**The results of model m.sub are an Adjusted R-squared value of 0.8175, an F-statistic of 56.45, and a highly significant p-value of <2.2e-16.**

The subset Age:Taxes is not significant, so it can be trimmed from the model.

m.sub1 <- lm(RentRate ~ Age+OperExp+W2MiDT+Age:OperExp+OperExp:W2MiDT+SqFt:W2MiDT, data=comm\_prop)  
summary(m.sub1)

##   
## Call:  
## lm(formula = RentRate ~ Age + OperExp + W2MiDT + Age:OperExp +   
## OperExp:W2MiDT + SqFt:W2MiDT, data = comm\_prop)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.02258 -0.45533 0.02524 0.50456 2.58106   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.247e+01 5.700e-01 21.883 < 2e-16 \*\*\*  
## Age -2.551e-01 5.638e-02 -4.525 1.80e-05 \*\*\*  
## OperExp 2.655e-01 6.584e-02 4.032 0.000114 \*\*\*  
## W2MiDTYes 1.926e+00 7.214e-01 2.669 0.008981 \*\*   
## Age:OperExp 1.577e-02 5.845e-03 2.698 0.008303 \*\*   
## OperExp:W2MiDTYes -2.262e-01 7.520e-02 -3.008 0.003389 \*\*   
## W2MiDTNo:SqFt 1.306e-06 1.208e-06 1.082 0.282278   
## W2MiDTYes:SqFt 1.112e-05 1.482e-06 7.503 3.84e-11 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.8135 on 92 degrees of freedom  
## Multiple R-squared: 0.83, Adjusted R-squared: 0.8171   
## F-statistic: 64.18 on 7 and 92 DF, p-value: < 2.2e-16

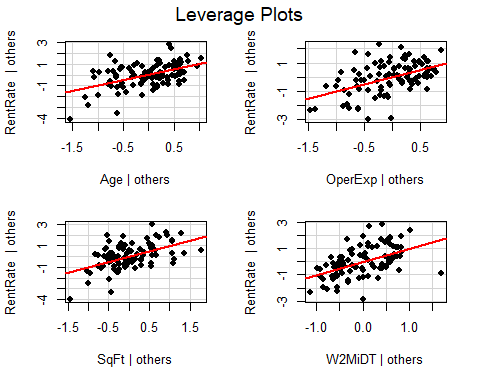
**The Adjusted R-squared value decreased slightly to 0.8171. The F-Statistic increased to 64.18 and the p-value had no change.**

## Review and Interpretation of diagnostic plots

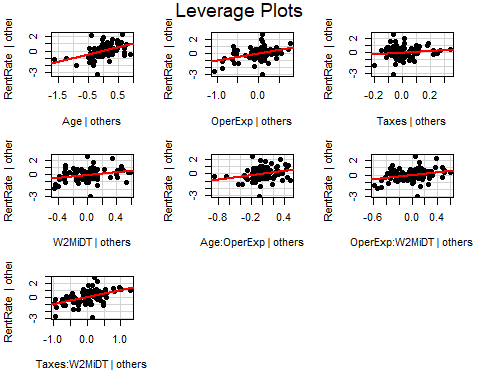
<strong>Adjusted R-sqrd F-Statistic</strong>   
 m.2 0.7313 68.35   
 m.step 0.8171 64.18  
 m.sub1 0.8171 64.18

### Models m.step1 and m.sub1 appear to to have the same values for F-Statistic and Adjusted R-squared the most appropriate predictive model. Further examination will need to be conducted to determine the most appropriate model.

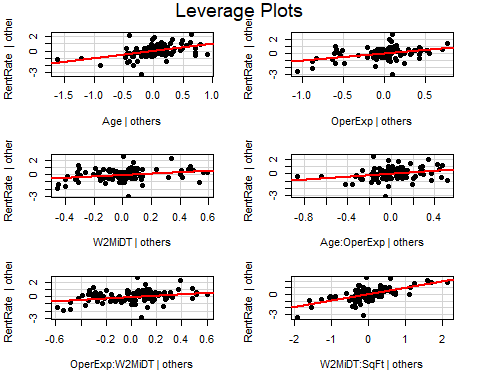
leveragePlots(m.1,pch=16)



leveragePlots(m.step,pch=16)



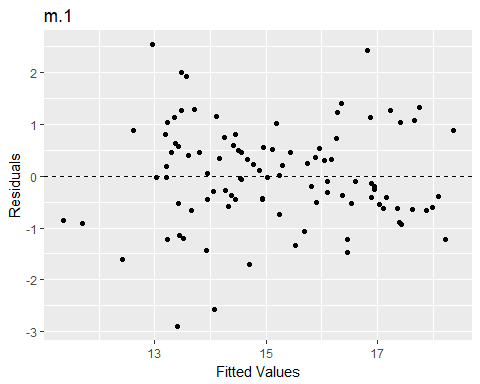
leveragePlots(m.sub1,pch=16)



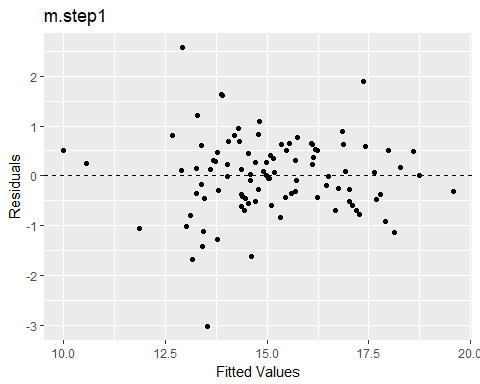
Observations: Leverage Plots are a way to visualize which variables are significant. â If the red line is not close to the horizontal line at y = 0, the variable is considered to be significant

### Residuals Variance Plot

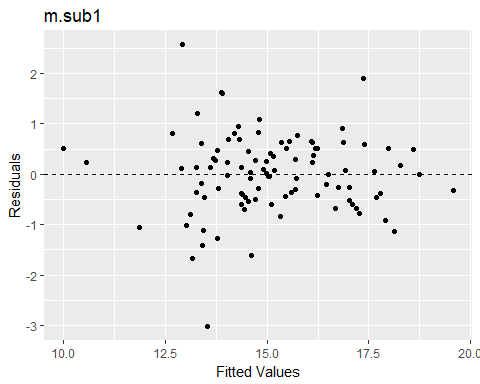
library(ggplot2)  
m\_df <- fortify(m.1)  
ggplot(m\_df, aes(x=.fitted, y=.resid)) +  
 geom\_point() +  
 geom\_hline(yintercept = 0, linetype = 2) +  
 labs(x="Fitted Values", y="Residuals", title="m.1")



m.step\_df <- fortify(m.step)  
ggplot(m.step\_df, aes(x=.fitted, y=.resid)) +  
 geom\_point() +  
 geom\_hline(yintercept = 0, linetype = 2) +  
 labs(x="Fitted Values", y="Residuals", title="m.step1")



m.sub\_df <- fortify(m.sub1)  
ggplot(m.sub\_df, aes(x=.fitted, y=.resid)) +  
 geom\_point() +  
 geom\_hline(yintercept = 0, linetype = 2) +  
 labs(x="Fitted Values", y="Residuals", title="m.sub1")



Observations: Models m.step and m.sub1 both seem to cluster in the middle section of the plot. This suggests non-constant variance.

### Checking for constant variance

car::ncvTest(m.1)

## Non-constant Variance Score Test   
## Variance formula: ~ fitted.values   
## Chisquare = 3.279042 Df = 1 p = 0.07016989

car::ncvTest(m.step)

## Non-constant Variance Score Test   
## Variance formula: ~ fitted.values   
## Chisquare = 8.191176 Df = 1 p = 0.004209462

car::ncvTest(m.sub1)

## Non-constant Variance Score Test   
## Variance formula: ~ fitted.values   
## Chisquare = 8.186369 Df = 1 p = 0.004220632

Conclusion: Significantly low p-values for the ncvTest confirm that both the m.step and m.sub1 models have non-constant variance. The m.1 model maintains constant variance, however.

### Checking for Normality of Residuals

shapiro.test(m.1$residuals)

##   
## Shapiro-Wilk normality test  
##   
## data: m.1$residuals  
## W = 0.98932, p-value = 0.6099

shapiro.test(m.step$residuals)

##   
## Shapiro-Wilk normality test  
##   
## data: m.step$residuals  
## W = 0.96971, p-value = 0.021

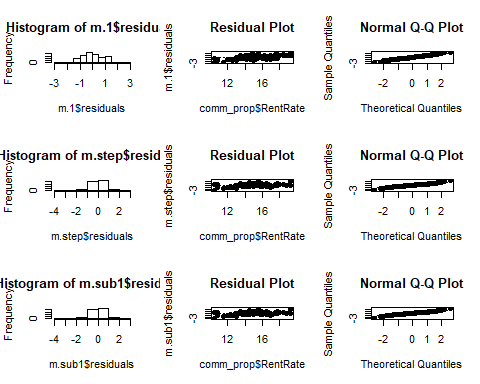
shapiro.test(m.sub1$residuals)

##   
## Shapiro-Wilk normality test  
##   
## data: m.sub1$residuals  
## W = 0.96969, p-value = 0.02092

Conclusion: Significantly low p-values for the Shapiro-Wilk test confirm that both the m.step and m.sub1 models’ are not normally distributed. The m.1 model is normal.

### Visual Representation of Normality and Constant Variance Tests

par(mfrow=c(3,3))  
  
hist(m.1$residuals)  
plot(m.1$residuals ~ comm\_prop$RentRate,pch=16, main="Residual Plot")  
abline(h = 0, lty = 3) # adds a horizontal dashed line at y = 0  
qqnorm(m.1$residuals, pch=16)  
qqline(m.1$residuals)  
  
hist(m.step$residuals)  
plot(m.step$residuals ~ comm\_prop$RentRate,pch=16, main="Residual Plot")  
abline(h = 0, lty = 3) # adds a horizontal dashed line at y = 0  
qqnorm(m.step$residuals, pch=16)  
qqline(m.step$residuals)  
  
hist(m.sub1$residuals)  
plot(m.sub1$residuals ~ comm\_prop$RentRate,pch=16, main="Residual Plot")  
abline(h = 0, lty = 3) # adds a horizontal dashed line at y = 0  
qqnorm(m.sub1$residuals, pch=16)  
qqline(m.sub1$residuals)



### Test for Multicolinearity

car::vif(m.1)

## Age OperExp SqFt W2MiDT   
## 1.459694 1.366874 1.582381 1.527237

car::vif(m.step)

## Age OperExp Taxes W2MiDT Age:OperExp   
## 20.061435 4.484482 2.744903 18.120465 26.238424   
## OperExp:W2MiDT Taxes:W2MiDT   
## 23.677754 7.479903

car::vif(m.sub1)

## GVIF Df GVIF^(1/(2\*Df))  
## Age 20.061044 1 4.478956  
## OperExp 4.484811 1 2.117737  
## W2MiDT 18.120204 1 4.256783  
## Age:OperExp 26.237538 1 5.122259  
## OperExp:W2MiDT 23.682733 1 4.866491  
## W2MiDT:SqFt 8.618614 2 1.713402

Conclusion: Tests suggest no Multicolinearity.

**Without further data transformation, due to the non-constant variance and disnormality of the m.step and m.sub1 models, m.1 is the most appropriate model to use in our prediction.**

new\_obs = data.frame(Age=9, SqFt=40,OperExp=13,W2MiDT="No")  
predict(m.1, newdata = new\_obs, interval="predict")

## fit lwr upr  
## 1 14.19777 12.16037 16.23517

new\_obs = data.frame(Age=9, SqFt=40,OperExp=13,Taxes=.54,W2MiDT="No")  
predict(m.step, newdata = new\_obs, interval="predict")

## fit lwr upr  
## 1 15.51173 13.80104 17.22242

new\_obs = data.frame(Age=9, SqFt=40,OperExp=13,Taxes=.54,W2MiDT="No")  
predict(m.sub1, newdata = new\_obs, interval="predict")

## fit lwr upr  
## 1 15.47277 13.74354 17.20199

**Prediction: We are 95% confidence that the value of RentRate will be between $12,120.37 and $16,235.17. With the combined costs of Operating Expenses and Taxes, $13540.00, we cannot say with confidence that we will be able to “offset the costs of ownership.” We should sell the building.**