MLR in prediciting House Prices in Toronto and Missiauga

TONGFEI ZHOU, Id 1004738448

December 1, 2020

I. Data Wrangling

First we sample 150 points randomly from our dataset, with seed 1004738448 as my student number, and then I show the IDs of the sample selected.

(a)

```
set.seed(1004738448)
data_orig_TZ <- read.csv("real203.csv")
data_raw_8448 <- sample(sample_n(data_orig_TZ, 150, replace=FALSE))
data_raw_8448$ID
```

```
##   [1]  68 135 195 187   2  67 171  97  76 186  50  48  96 109 207  82  69  95
##  [19]  84 163  62 205 177  60  42 185  36 155  90 145  33 117  24 178  26 180
##  [37]  30  91  11  40  49 201 152 143  85 115 151 191 196   7 194 183 138  86
##  [55] 111 142 107 149 132 181 169  21  43   5  14 108  75 139 175 103  23 172
##  [73]  27  58  46 193  57 165 104 176   6 162 141   4 166  63 112  39 137  70
##  [91]  78  10 101  64 164 173  19  83  77  22 146  17 188 148 218  92 189  15
## [109]  53 227  45 147 102 204  35   3  56 174 157  81  54 160  98  28  94 110
## [127] 144 182 119 168 154 114 106 167 122 105 140  74  44 131  65  93  80  61
## [145] 134 190  55 170  72 125
```

So the IDs are reported above.

(b) Now we create new variable lotsize = lotwidth * lotlength and replace lotwidth and lotlength.

```
dat_8448 <- data_raw_8448 %>% mutate(lotsize = lotlength * lotwidth) %>%
  select(-c(lotlength, lotwidth))
```

(c) Next we clean our data. As one can see from the summary list, there are 92 missing values from variable maxsqfoot, and so this is very bad and I remove this predictor away. What left over are total 8 missing values among taxes, parking ad lotsize, so I remove these data points to help the analysis below easier.

```
summary(dat_8448)
```

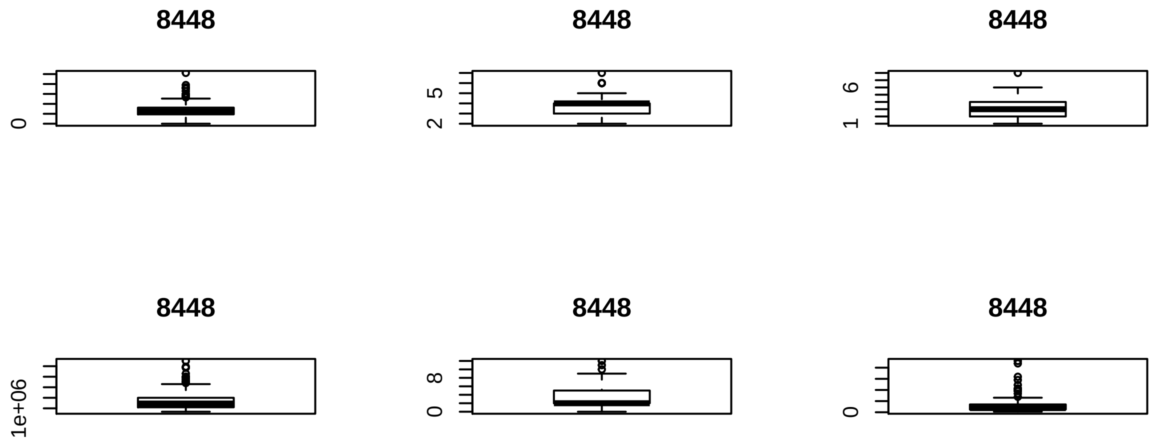
```
##       sale                ID           maxsqfoot     location     taxes
##  Min.   : 672000   Min.   :  2.00   Min.   :1500   M:61     Min.   :    4.375
```

```
##  1st Qu.:1106250   1st Qu.: 57.25   1st Qu.:2000   T:89     1st Qu.: 4527.000
##  Median :1400000   Median :104.50   Median :2500            Median : 6083.000
##  Mean   :1729070   Mean   :106.71   Mean   :2922            Mean   : 7175.196
##  3rd Qu.:2104500   3rd Qu.:161.50   3rd Qu.:3500            3rd Qu.: 8040.000
##  Max.   :5100000   Max.   :227.00   Max.   :5000            Max.   :25575.000
##                                     NA's   :92              NA's   :1
##       list            parking          bedroom         bathroom
##  Min.   : 649000   Min.   : 0.00   Min.   :1.000   Min.   :1.0
##  1st Qu.:1011500   1st Qu.: 2.00   1st Qu.:3.000   1st Qu.:2.0
##  Median :1424000   Median : 2.00   Median :4.000   Median :3.0
##  Mean   :1736498   Mean   : 3.34   Mean   :3.613   Mean   :3.3
##  3rd Qu.:1999000   3rd Qu.: 4.25   3rd Qu.:4.000   3rd Qu.:4.0
##  Max.   :5499000   Max.   :12.00   Max.   :7.000   Max.   :8.0
##                    NA's   :6
##     lotsize
##  Min.   :  297.4
##  1st Qu.: 2441.2
##  Median : 3599.0
##  Mean   : 6050.4
##  3rd Qu.: 6766.0
##  Max.   :46057.3
##  NA's   :1
```

```r
dat_8448 <- dat_8448 %>% select(-maxsqfoot)
data_TZ <- na.omit(dat_8448)
```

Further, we check the potential serious outliers by boxplots, and one can see that there is one point in taxes variable and there are two points in lotsize variable that have quite extreme value, and so I identify them and found that they are just great mansions with 10+ parking pots and 5+ bedrooms and bathrooms, and so I will remove them since they are too glorious for normal houses. Therefore, we have 139 obs and 9 variables(included IDs) after all data cleaning.

```r
par(mfrow = c(3,3))
boxplot(data_TZ$taxes, main="8448")
boxplot(data_TZ$bedroom, main="8448")
boxplot(data_TZ$bathroom, main="8448")
boxplot(data_TZ$list, main="8448")
boxplot(data_TZ$parking, main="8448")
boxplot(data_TZ$lotsize, main="8448")
max_taxes <- max(data_TZ$taxes)
second_max_lotsize <- sort(data_TZ$lotsize, decreasing = T)[2]
data_TZ <- filter(data_TZ, data_TZ$taxes < max_taxes)
data_TZ <- filter(data_TZ, data_TZ$lotsize < second_max_lotsize)
```



II. Exploratory Data Analysis

```
str(data_TZ)
```

```
## 'data.frame':    139 obs. of  9 variables:
##  $ sale    : int  1128000 1450000 2270000 1625000 2200000 1410000 930000 1550000 1140000 3300000 ...
##  $ ID      : int  68 135 195 187 2 67 171 97 186 50 ...
##  $ location: Factor w/ 2 levels "M","T": 2 1 1 1 2 2 1 2 1 2 ...
##  $ taxes   : num  4494 6484 12200 7687 7712 ...
##  $ list    : int  1149000 1484000 2300000 1639500 1999900 1449000 939000 1588000 1199000 3595000 ...
##  $ parking : int  1 6 7 4 3 2 4 1 6 2 ...
##  $ bedroom : int  3 4 5 4 5 3 4 4 4 5 ...
##  $ bathroom: int  2 3 4 4 3 2 4 4 4 5 ...
##  $ lotsize : num  2139 5729 21300 6000 4664 ...
##  - attr(*, "na.action")= 'omit' Named int  9 13 14 19 120 132 144 147
##   ..- attr(*, "names")= chr  "9" "13" "14" "19" ...
```

(a) From the structure output above, we can see that the categorical variable is location, here it is stored as a factor type. The discrete variables are number of parking, the number of bedrooms, the number of bathrooms and IDs. The continuous variables are sale, taxes, list and lotsize.

(b) Below are pairwise correlations and scatterplot matrix for all pairs of quantitative variables (notice that in data analysis we will not consider IDs anymore since it is only served as an identification use).

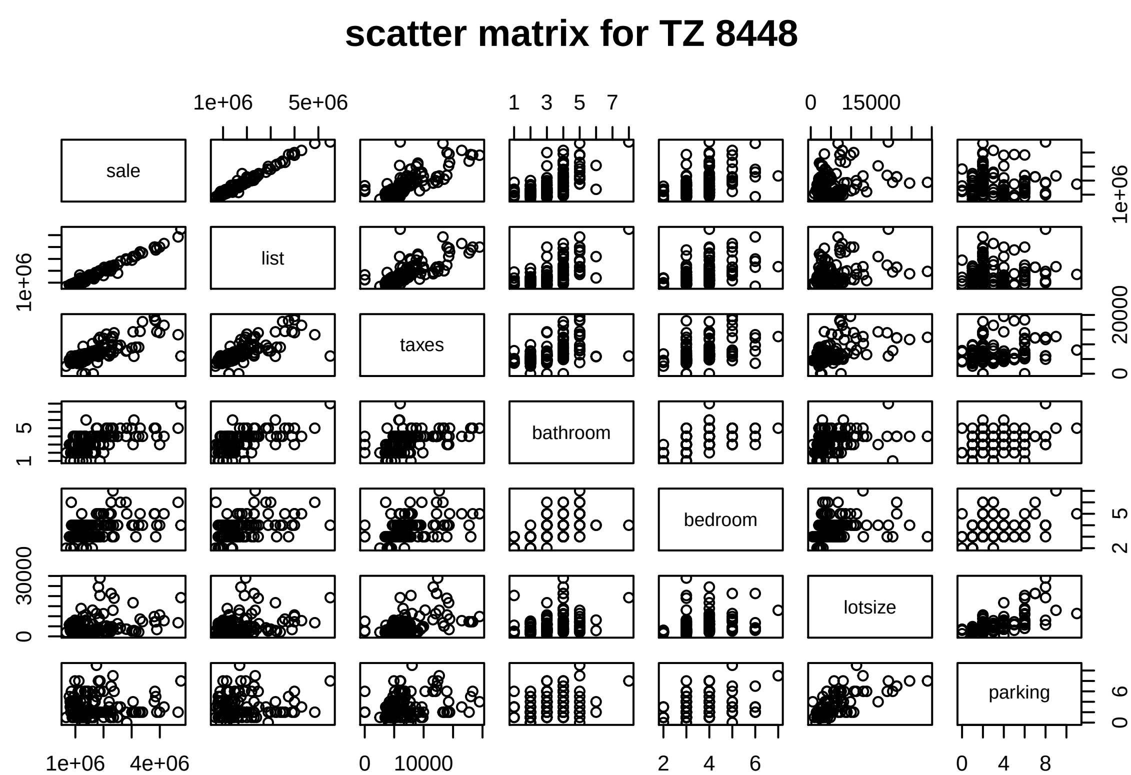
```r
data_TZ <- data_TZ %>% mutate(location = as.numeric(location) - 1)
attach(data_TZ)
numericx <- cbind(sale, list, bedroom, bathroom, taxes, parking, lotsize)
cor_matrix <- round(cor(numericx), 4)
cor_matrix
```

```
##            sale   list bedroom bathroom  taxes parking lotsize
## sale     1.0000 0.9871  0.4634   0.5936 0.7646  0.0149  0.2799
## list     0.9871 1.0000  0.4660   0.6080 0.7493  0.0508  0.3041
## bedroom  0.4634 0.4660  1.0000   0.5655 0.4387  0.3241  0.2799
## bathroom 0.5936 0.6080  0.5655   1.0000 0.4829  0.3064  0.3129
## taxes    0.7646 0.7493  0.4387   0.4829 1.0000  0.2258  0.4811
## parking  0.0149 0.0508  0.3241   0.3064 0.2258  1.0000  0.6842
## lotsize  0.2799 0.3041  0.2799   0.3129 0.4811  0.6842  1.0000
```

So for sale price rank in terms of the correlation coefficients, the predictors from the highest to lowest are the following: list, taxes, bathroom, bedroom, lotsize, parking.

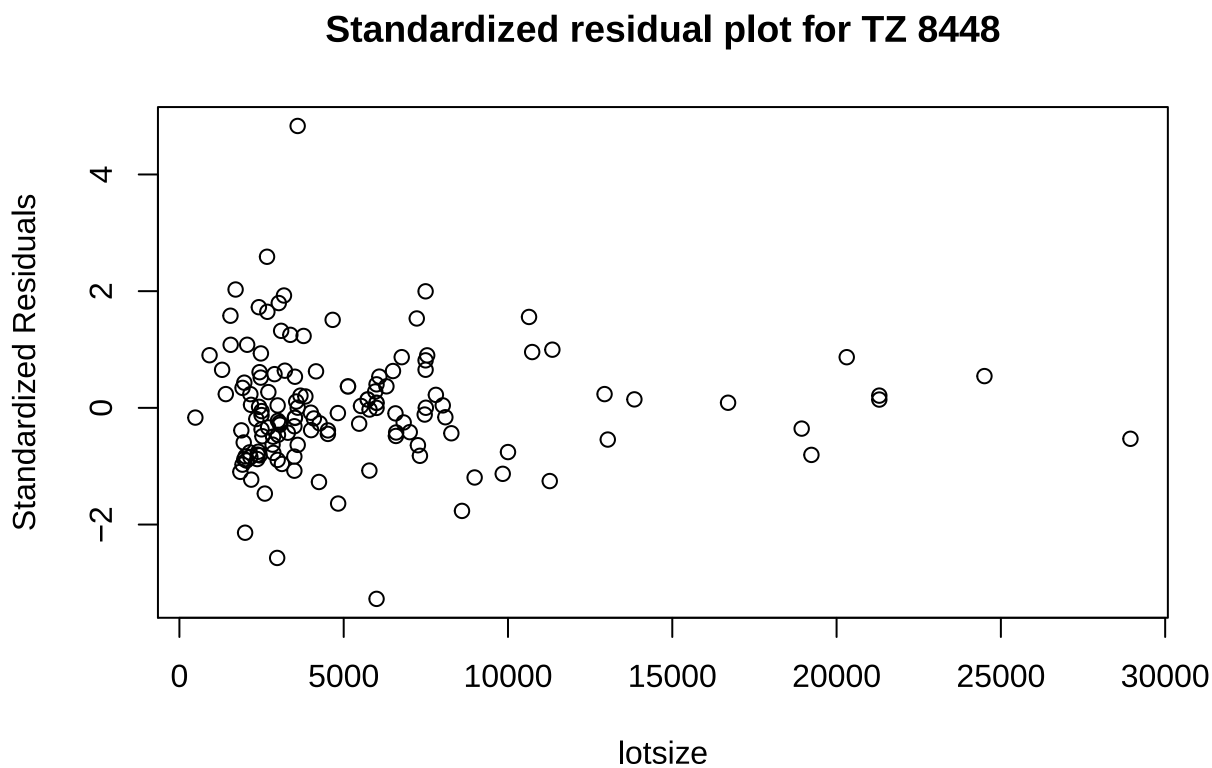
Now is the scatterplot matrix.

```r
pairs(sale~list+taxes+bathroom+bedroom+lotsize+parking, data = data_TZ, cex.labels = 0.85, main="scatte
```



(c) From the plot above, one can see that only the predictor lotsize has the potential violation of the assumption of constant variance since at the beginning of the value of lotsize, the variance of y is quite big, but it tends to decrease as the value of lotsize increases. Also we confirm this by showing the plot of standardized residuals against lotsize. (Here I changed the location from a factor variable into a numeric catergorical variable with values 0 and 1 when I attach the data variables so that it becomes a dummy variable instead of a factor)

```
lmod_full_8448 <- lm(sale~list+taxes+bathroom+bedroom+lotsize+parking+location, data = data_TZ)
stdres_8448 <- rstandard(lmod_full_8448)
plot(lotsize, stdres_8448, ylab = "Standardized Residuals", main = "Standardized residual plot for TZ 84
```



So as one can see from above, there is a "fanning in" pattern as lotsize increases. Therefore, violation of constant variance with predictor lotsize of sale price exists.

III. Methods and Model

(i) Now we fit an additive linear regression model with all available predictors, but actually we have done it above with lmod_full_8448. So here will be the list output.

```
kable(summary(lmod_full_8448)$coefficients,digits = 4, caption = "coefficents result for TZ 8448")
```

Table 1: coefficents result for TZ 8448

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	35097.9650	54935.3861	0.6389	0.5240
list	0.8350	0.0237	35.2441	0.0000

	Estimate	Std. Error	t value	Pr(>|t|)
taxes	20.6466	4.9910	4.1367	0.0001
bathroom	12237.0753	13585.5058	0.9007	0.3694
bedroom	8441.6157	15172.6918	0.5564	0.5789
lotsize	0.0474	3.5908	0.0132	0.9895
parking	-7803.3096	8710.3891	-0.8959	0.3720
location	93812.1011	37374.4467	2.5101	0.0133

So here I report p-values to 4 decimal places and at significance level of $\alpha = 0.05$ as required, and we see that from above the p-value for list and taxes are strongly significant and location is somewhat significant.

Therefore, the interpretation for list is that holding other predictors unchanged, with one dollar increasing in the last list price of the property, the actual sale price of the property will on average increase 0.835 dollar. For taxes, it means that holding other predictors constant, with one dollar increasing in the previous year's property tax, the average actual sale price of the property will increase 20.6466 dollar.

Finally, for the variable location, it means that on average speaking, when all other variables are constant, sale price differs at Toronto or Missiagua, at an average level of 93812.1011 dollar, so buying house in Toronto is much more expensive than it is in Missiagua.

(ii) Now we start with the full model obtained above, and use backward elimination with AIC.

```r
step(lmod_full_8448, direction = "backward")
```

```
## Start:  AIC=3272.82
## sale ~ list + taxes + bathroom + bedroom + lotsize + parking +
##     location
##
##              Df  Sum of Sq        RSS    AIC
## - lotsize     1 2.7678e+06 2.0830e+12 3270.8
## - bedroom     1 4.9220e+09 2.0879e+12 3271.1
## - parking     1 1.2761e+10 2.0957e+12 3271.7
## - bathroom    1 1.2901e+10 2.0959e+12 3271.7
## <none>                     2.0830e+12 3272.8
## - location    1 1.0018e+11 2.1832e+12 3277.3
## - taxes       1 2.7210e+11 2.3551e+12 3287.9
## - list        1 1.9751e+13 2.1834e+13 3597.4
##
## Step:  AIC=3270.82
## sale ~ list + taxes + bathroom + bedroom + parking + location
##
##              Df  Sum of Sq        RSS    AIC
## - bedroom     1 4.9308e+09 2.0879e+12 3269.1
## - bathroom    1 1.3080e+10 2.0961e+12 3269.7
## - parking     1 1.5521e+10 2.0985e+12 3269.9
## <none>                     2.0830e+12 3270.8
## - location    1 1.0690e+11 2.1899e+12 3275.8
## - taxes       1 2.9288e+11 2.3759e+12 3287.1
## - list        1 2.0584e+13 2.2667e+13 3600.6
##
## Step:  AIC=3269.15
## sale ~ list + taxes + bathroom + parking + location
##
```

```
##            Df  Sum of Sq        RSS    AIC
## - parking   1 1.2198e+10 2.1001e+12 3268.0
## - bathroom  1 2.2089e+10 2.1100e+12 3268.6
## <none>                   2.0879e+12 3269.1
## - location  1 1.1700e+11 2.2049e+12 3274.7
## - taxes     1 3.0492e+11 2.3928e+12 3286.1
## - list      1 2.0624e+13 2.2712e+13 3598.9
##
## Step:  AIC=3267.96
## sale ~ list + taxes + bathroom + location
##
##            Df  Sum of Sq        RSS    AIC
## - bathroom  1 1.9640e+10 2.1198e+12 3267.3
## <none>                   2.1001e+12 3268.0
## - taxes     1 2.9662e+11 2.3967e+12 3284.3
## - location  1 3.3127e+11 2.4314e+12 3286.3
## - list      1 2.0749e+13 2.2849e+13 3597.7
##
## Step:  AIC=3267.25
## sale ~ list + taxes + location
##
##            Df  Sum of Sq        RSS    AIC
## <none>                   2.1198e+12 3267.3
## - taxes     1 2.8601e+11 2.4058e+12 3282.8
## - location  1 3.2220e+11 2.4420e+12 3284.9
## - list      1 3.0477e+13 3.2597e+13 3645.1
```

```
##
## Call:
## lm(formula = sale ~ list + taxes + location, data = data_TZ)
##
## Coefficients:
## (Intercept)         list        taxes     location
##   5.690e+04    8.472e-01    2.015e+01    1.066e+05
```

So the final fitted model is

$$\hat{Y}_{i,sale} = \hat{\beta}_0 + \hat{\beta}_1 X_{i,list} + \hat{\beta}_2 X_{i,taxes} + \hat{\beta}_3 X_{i,location}$$

This is consistent with what we concluded in part (i), since in part (i) these three variables are which βs' are statistically significant.

(iii) Now we use BIC.

```
n <- length(sale)
step(lmod_full_8448, direction = "backward", k=log(n))
```

```
## Start:  AIC=3296.29
## sale ~ list + taxes + bathroom + bedroom + lotsize + parking +
##     location
##
##            Df  Sum of Sq        RSS    AIC
## - lotsize   1 2.7678e+06 2.0830e+12 3291.4
```

```
## - bedroom   1 4.9220e+09 2.0879e+12 3291.7
## - parking   1 1.2761e+10 2.0957e+12 3292.2
## - bathroom  1 1.2901e+10 2.0959e+12 3292.2
## <none>                  2.0830e+12 3296.3
## - location  1 1.0018e+11 2.1832e+12 3297.9
## - taxes     1 2.7210e+11 2.3551e+12 3308.4
## - list      1 1.9751e+13 2.1834e+13 3618.0
##
## Step:  AIC=3291.36
## sale ~ list + taxes + bathroom + bedroom + parking + location
##
##            Df  Sum of Sq        RSS    AIC
## - bedroom   1 4.9308e+09 2.0879e+12 3286.8
## - bathroom  1 1.3080e+10 2.0961e+12 3287.3
## - parking   1 1.5521e+10 2.0985e+12 3287.5
## <none>                  2.0830e+12 3291.4
## - location  1 1.0690e+11 2.1899e+12 3293.4
## - taxes     1 2.9288e+11 2.3759e+12 3304.7
## - list      1 2.0584e+13 2.2667e+13 3618.2
##
## Step:  AIC=3286.75
## sale ~ list + taxes + bathroom + parking + location
##
##            Df  Sum of Sq        RSS    AIC
## - parking   1 1.2198e+10 2.1001e+12 3282.6
## - bathroom  1 2.2089e+10 2.1100e+12 3283.3
## <none>                  2.0879e+12 3286.8
## - location  1 1.1700e+11 2.2049e+12 3289.4
## - taxes     1 3.0492e+11 2.3928e+12 3300.8
## - list      1 2.0624e+13 2.2712e+13 3613.6
##
## Step:  AIC=3282.63
## sale ~ list + taxes + bathroom + location
##
##            Df  Sum of Sq        RSS    AIC
## - bathroom  1 1.9640e+10 2.1198e+12 3279.0
## <none>                  2.1001e+12 3282.6
## - taxes     1 2.9662e+11 2.3967e+12 3296.1
## - location  1 3.3127e+11 2.4314e+12 3298.1
## - list      1 2.0749e+13 2.2849e+13 3609.5
##
## Step:  AIC=3278.99
## sale ~ list + taxes + location
##
##            Df  Sum of Sq        RSS    AIC
## <none>                  2.1198e+12 3279.0
## - taxes     1 2.8601e+11 2.4058e+12 3291.6
## - location  1 3.2220e+11 2.4420e+12 3293.7
## - list      1 3.0477e+13 3.2597e+13 3653.9
```

```
##
## Call:
## lm(formula = sale ~ list + taxes + location, data = data_TZ)
##
```

```
## Coefficients:
## (Intercept)         list        taxes     location
##   5.690e+04    8.472e-01    2.015e+01    1.066e+05
```

As we can from above, the final fitted model is still

$$\hat{Y}_{i,sale} = \hat{\beta}_0 + \hat{\beta}_1 X_{i,list} + \hat{\beta}_2 X_{i,taxes} + \hat{\beta}_3 X_{i,location}$$

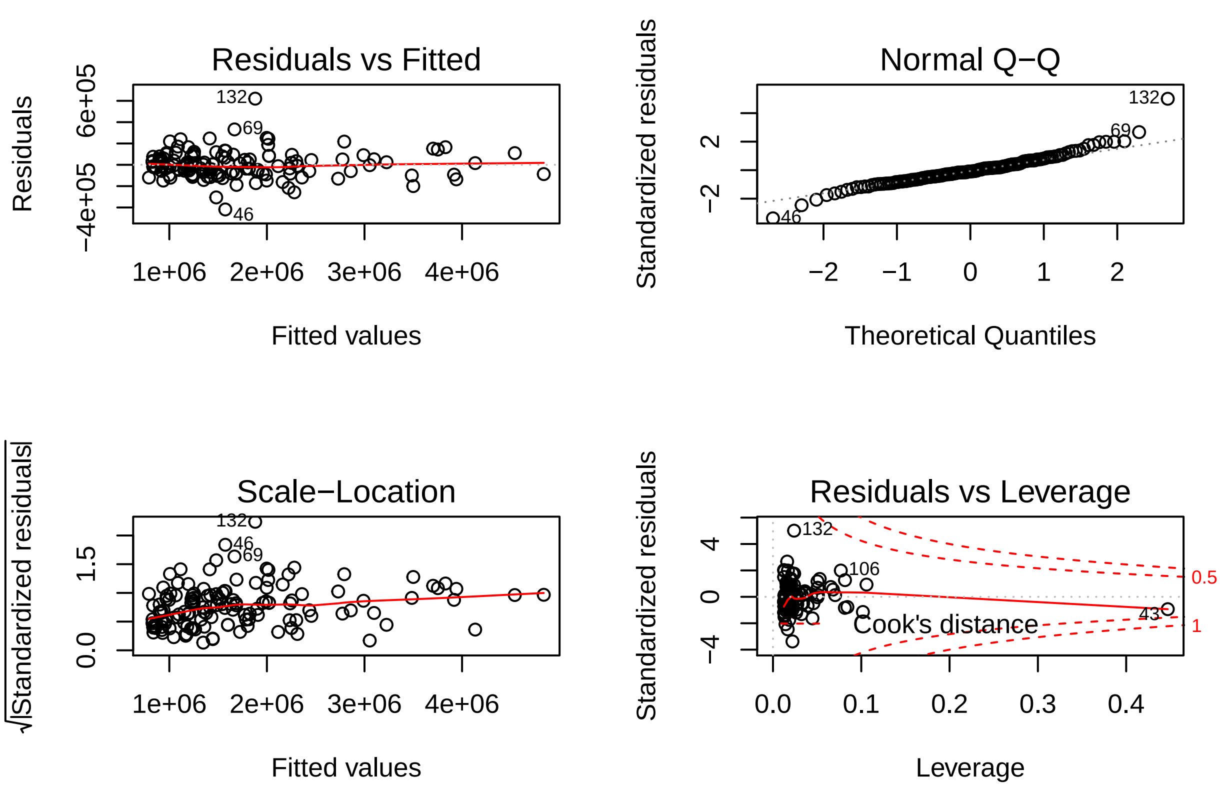
Therefore, the results are consistent with part (ii) and (i), for each step the elimination of variable is same with what has been done in part (ii).

Explanation is that since in both AIC and BIC, we are using the same elimination method, we will tend to contain the lowest p-value predictors in the full model, which is also the idea in part (i). Also, since here $n = 139 >> 9 = p$, and in R, $BIC = nlog(\frac{RSS}{n}) + plog(n)$ and $AIC = nlog(\frac{RSS}{n}) + 2p$, we notice that $log(139) \approx 5$ and so the difference in AIC and BIC here is quite small even though right now the penalty term in BIC is greater than what in AIC, but it is with only $3 \times 9 = 27$, which is extremely small compared to $nlog(\frac{RSS}{n})$. Therefore, the criterion AIC and BIC are almost same. As a consequence, these two approaches(AIC with backward elimination and BIC with backward elimination) will result in same results in this case.

IV. Discussions and Limitations

(a) Now we show the 4 diagnostic plots from the final model obtained from partIII (iii).

```r
par(mfrow = c(2,2))
lmod_final_8448 <- lm(sale~list+taxes+location, data = data_TZ)
plot(lmod_final_8448)
```



(b) As we can see from above, the line in Residuals vs. Fitted is almost straight, and no pattern is found here with all points scattered randomly around the horizontal zero line. For the plot of the square root of the absolute value of standardized residuals vs. fitted values, we see that all point scattered randomly without any pattern, and also there is no upward or downward curves,i.e, no trend existed. For the plot of standardized residuals vs. leverage, we can notice that all points are inside red lines, which means that no outliers or influential points. Finally, as for the Normal Q-Q plot, we can see that almost all points lie on the 45 degreee line, with only two points being far away. This shows that, fortunately, normality holds for this model. Therefore, our normal error MLR assumptions are all being satisfied well.(linear relationship, independence of error terms, normality and constant variance)

(c) For now we use AIC and BIC with backward elimination to select the variables, we can continue to check whether there are some predictors that are outside our analysis. For instance, the number of markets around the property, the year of use of the property, etc. Besides, we can continue to investigate whether an interaction term of the dummy variable "location" with some other predictor needs adding. Also, we could see if the model would be further reduced using partial F-test.