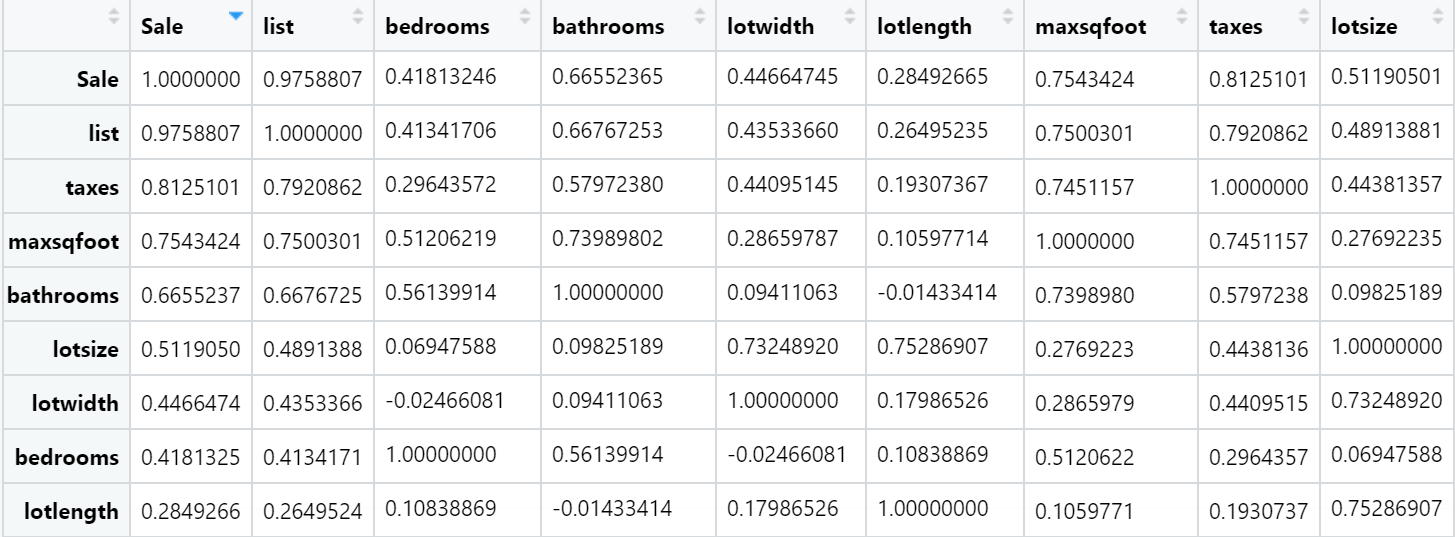
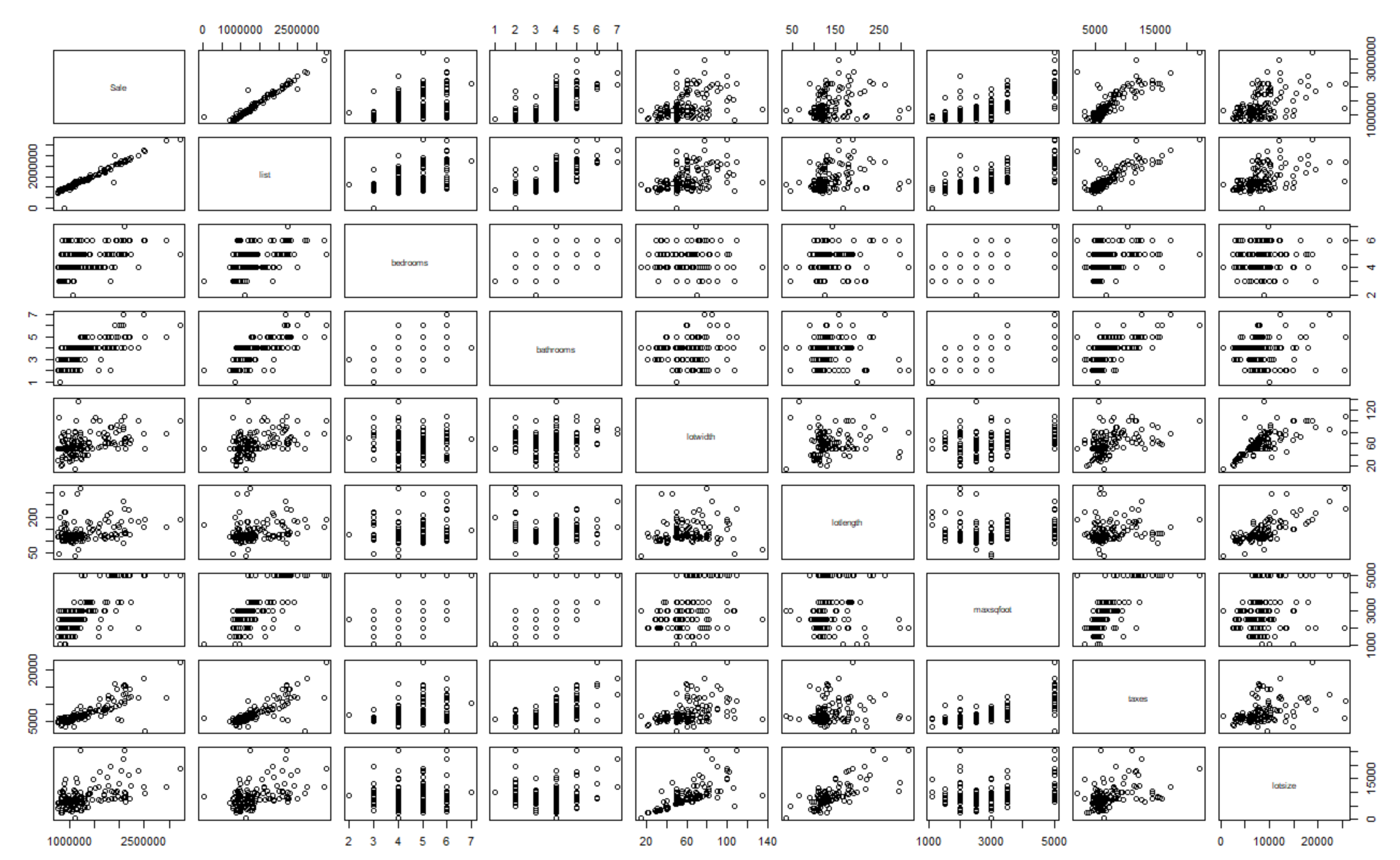
Real estate prices in GTA

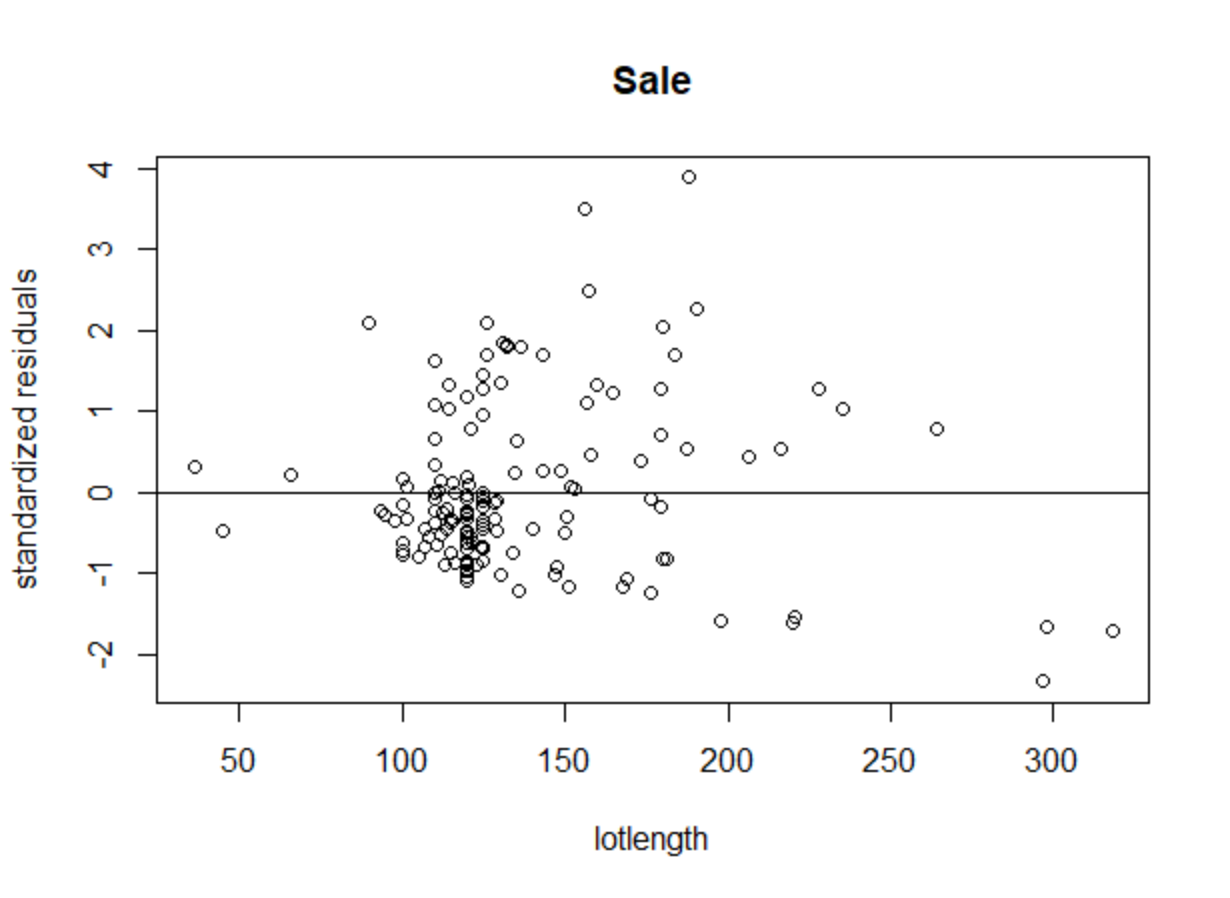
1.



From the pairwise correlations matrix, the rank of the quantitative predictors for sale price are:

list> taxes> maxsqfoot> bathrooms> lotsize> lotwidth> bedrooms> lotlength

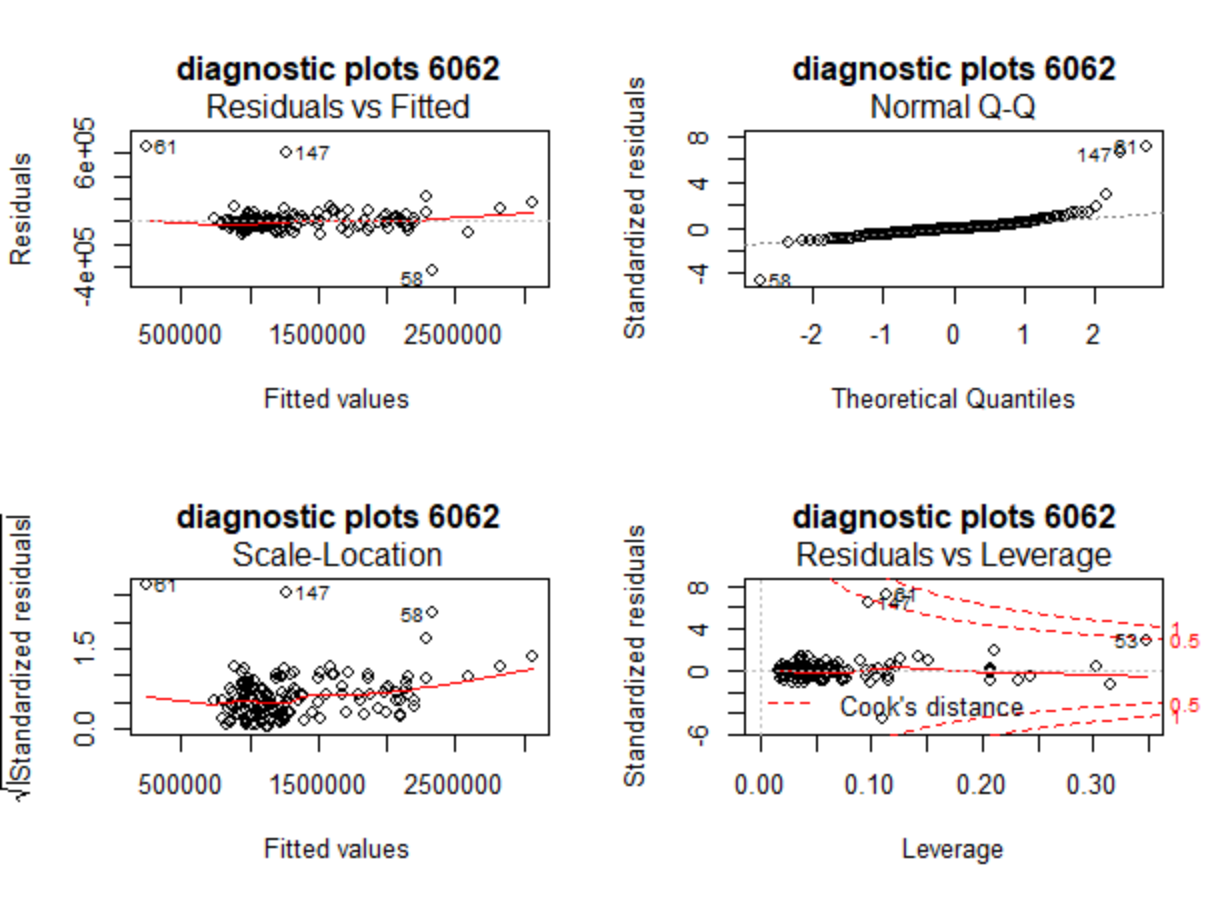
2. From observing the scatterplot matrix in part1, predictor lotlength seems to be violating the assumption of constant variance. A plot of the standardized residuals shows unbalanced distribution along the y-axis, and data are concentrated near (100,0) to (150,0), confirming that lotlength variable does violate the constant variance assumption. A possible solution is to transform predictor or the response variable.



3. The fitted multiple linear regression model is Sale=67790+0.7568list+9108bedrooms+9594bathrooms-409.2lotwidth-111.9lotlength+3.898maxsqfoot+14.50taxes-1086locationX+7.988lotsize

|  |  |  |
| --- | --- | --- |
|  | Estimated regression coefficients | p-values |
| **list** | **0.7568** | **<2.2e-16(0)** |
| bedrooms | 9108 | 0.3865 |
| bathrooms | 9594 | 0.4366 |
| lotwidth | -409.2 | 0.7004 |
| lotlength | -111.9 | 0.8383 |
| maxsqfoot | 3.898 | 0.7862 |
| **taxes** | **14.50** | **0.0024(0.001)** |
| locationX | -1086 | 0.9464 |
| lotsize | 7.988 | 0.3098 |

With a significance level of 5%, only variables list and taxes are significant. For every dollar increase in list price, the average sale price is $0.7568 higher. For every dollar increase in taxes, the average sale price is $14.50 higher.

4. 

I consider Case\_ID 61 as an influential point. To find points with high leverage (high hii), a threshold h\*ii = 0.1235 is calculated and the points with higher than threshold hii are found through sorting. After that, points with the largest Cook’s Distances are also found through sorting. My threshold for influential point is to have a Cook’s Distance>0.5, which leaves Case\_ID 61 with Cook’s Distance of 0.6633, and it is also consistent with the Residuals vs. Leverage graph.

5. The fitted model using backward elimination with AIC is Sale=71160+0.7609list+15910bathrooms+14.34taxes+5.750lotsize. The results are not consistent with those in part3 as this one ends with 4 significant predictors instead of 2.

6. Using BIC, the final model is Sale=116700+0.7957list+16.26taxes. The results are inconsistent with part 3 and 5 because from definition : and BIC will be bigger for larger n, which happens with this dataset as final BIC=3736.15>final AIC=3724.29. In general, AIC tends to give larger, more complex models while BIC gives smaller models since the penalty for model complexity is heavier.

**Appendix**

> setwd("~/sta302 codes")

> a3data <- read.table("reale\_a3data.csv", header = TRUE, sep = ",")

> a3data\_new <- na.omit(a3data)

> a3data\_new$lotsize = a3data\_new$lotwidth \* a3data\_new$lotlength

> attach(a3data\_new)

> set1 = subset(a3data\_new, select = -c(Case\_ID, location))

> set1.cor = cor(set1)

> pairs(set1, pch=1)

> sale.lm = lm(Sale~lotlength, data = set1)

> sale.stdres = rstandard(sale.lm)

> plot(lotlength, sale.stdres, ylab = "standardized residuals", xlab = "lotlength", main = "Sale")

> abline(0,0)

> dataq3 = subset(a3data\_new, select = -c(Case\_ID))

> m1 <- lm(Sale~list + bedrooms + bathrooms + lotwidth + lotlength + maxsqfoot + taxes + location + lotsize)

> summary(m1)

Call:

lm(formula = Sale ~ list + bedrooms + bathrooms + lotwidth +

lotlength + maxsqfoot + taxes + location + lotsize)

Residuals:

Min 1Q Median 3Q Max

-427339 -40483 -6274 19780 661411

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 6.779e+04 8.483e+04 0.799 0.42550

list 7.568e-01 2.962e-02 25.553 < 2e-16 \*\*\*

bedrooms 9.108e+03 1.049e+04 0.869 0.38648

bathrooms 9.594e+03 1.230e+04 0.780 0.43660

lotwidth -4.092e+02 1.061e+03 -0.386 0.70037

lotlength -1.119e+02 5.475e+02 -0.204 0.83829

maxsqfoot 3.898e+00 1.434e+01 0.272 0.78617

taxes 1.450e+01 4.685e+00 3.094 0.00235 \*\*

locationX -1.086e+03 1.613e+04 -0.067 0.94642

lotsize 7.988e+00 7.838e+00 1.019 0.30976

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 97890 on 152 degrees of freedom

Multiple R-squared: 0.9586, Adjusted R-squared: 0.9562

F-statistic: 391.2 on 9 and 152 DF, p-value: < 2.2e-16

> par(mfrow=c(2,2))

> plot(m1, main = "diagnostic plots 6062")

> pp1 = length(coef(m1))

> threshold=2\*(pp1)/length(Sale)

> threshold

[1] 0.1234568

> hii=hatvalues(m1)

> set1[hii>threshold, ]

Sale list bedrooms bathrooms lotwidth lotlength maxsqfoot taxes lotsize

8 860000 899000 6 3 35.01 296.50 2500 5740 10380.4650

13 3220000 3249850 5 6 100.00 188.00 5000 22244 18800.0000

20 1150000 1150000 5 2 45.67 297.87 2000 6421 13603.7229

22 800000 799900 3 2 106.90 45.00 3000 5686 4810.5000

27 1175000 1199000 4 4 135.63 66.00 2500 5584 8951.5800

41 2080000 2189000 6 7 85.00 264.00 5000 12652 22440.0000

48 2375000 2499000 4 4 100.00 179.83 2000 11768 17983.0000

53 1210000 1250000 4 2 80.08 318.04 2000 5843 25468.6432

55 2508000 2680000 6 5 50.07 190.40 5000 2073 9533.3280

64 2950000 3199900 6 5 77.00 156.20 5000 11745 12027.4000

99 1120000 1135000 4 4 14.02 36.06 3000 6476 505.5612

112 1821000 1830000 3 2 90.00 216.40 2000 8305 19476.0000

158 2100000 2199000 6 5 109.18 235.01 5000 11026 25658.3918

> round(sort(hii,decreasing=TRUE)[1:13],4)

53 7 25 155 51 11 20 96 39 18 46 62

0.3487 0.3164 0.3036 0.2439 0.2322 0.2113 0.2081 0.2074 0.2074 0.2070 0.1509 0.1421

109

0.1253

> cooks<-cooks.distance(m1)

> round(sort(cooks,decreasing=TRUE)[1:13],4)

61 147 53 58 11 7 51 62 39 109 46 136

0.6633 0.4622 0.4364 0.2628 0.0902 0.0706 0.0311 0.0296 0.0243 0.0171 0.0141 0.0137

55

0.0126

> step(m1,direction = "backward")

Start: AIC=3732.96

Sale ~ list + bedrooms + bathrooms + lotwidth + lotlength + maxsqfoot +

taxes + location + lotsize

Df Sum of Sq RSS AIC

- location 1 4.3414e+07 1.4566e+12 3731.0

- lotlength 1 4.0049e+08 1.4569e+12 3731.0

- maxsqfoot 1 7.0777e+08 1.4573e+12 3731.0

- lotwidth 1 1.4244e+09 1.4580e+12 3731.1

- bathrooms 1 5.8302e+09 1.4624e+12 3731.6

- bedrooms 1 7.2284e+09 1.4638e+12 3731.8

- lotsize 1 9.9528e+09 1.4665e+12 3732.1

<none> 1.4565e+12 3733.0

- taxes 1 9.1754e+10 1.5483e+12 3740.9

- list 1 6.2570e+12 7.7136e+12 4001.0

Step: AIC=3730.96

Sale ~ list + bedrooms + bathrooms + lotwidth + lotlength + maxsqfoot +

taxes + lotsize

Df Sum of Sq RSS AIC

- lotlength 1 3.7519e+08 1.4570e+12 3729.0

- maxsqfoot 1 6.6588e+08 1.4573e+12 3729.0

- lotwidth 1 1.4106e+09 1.4580e+12 3729.1

- bathrooms 1 5.9660e+09 1.4626e+12 3729.6

- bedrooms 1 7.1972e+09 1.4638e+12 3729.8

- lotsize 1 9.9282e+09 1.4665e+12 3730.1

<none> 1.4566e+12 3731.0

- taxes 1 9.3091e+10 1.5497e+12 3739.0

- list 1 6.2864e+12 7.7430e+12 3999.6

Step: AIC=3729

Sale ~ list + bedrooms + bathrooms + lotwidth + maxsqfoot + taxes +

lotsize

Df Sum of Sq RSS AIC

- maxsqfoot 1 6.0072e+08 1.4576e+12 3727.1

- lotwidth 1 1.4436e+09 1.4584e+12 3727.2

- bathrooms 1 6.7364e+09 1.4637e+12 3727.8

- bedrooms 1 6.9494e+09 1.4639e+12 3727.8

<none> 1.4570e+12 3729.0

- lotsize 1 4.4471e+10 1.5014e+12 3731.9

- taxes 1 9.4352e+10 1.5513e+12 3737.2

- list 1 6.2861e+12 7.7431e+12 3997.6

Step: AIC=3727.07

Sale ~ list + bedrooms + bathrooms + lotwidth + taxes + lotsize

Df Sum of Sq RSS AIC

- lotwidth 1 1.3002e+09 1.4589e+12 3725.2

- bedrooms 1 8.4027e+09 1.4660e+12 3726.0

- bathrooms 1 9.3962e+09 1.4670e+12 3726.1

<none> 1.4576e+12 3727.1

- lotsize 1 4.3912e+10 1.5015e+12 3729.9

- taxes 1 1.1766e+11 1.5752e+12 3737.6

- list 1 6.5242e+12 7.9817e+12 4000.5

Step: AIC=3725.22

Sale ~ list + bedrooms + bathrooms + taxes + lotsize

Df Sum of Sq RSS AIC

- bedrooms 1 9.6765e+09 1.4685e+12 3724.3

- bathrooms 1 9.7008e+09 1.4686e+12 3724.3

<none> 1.4589e+12 3725.2

- lotsize 1 5.6582e+10 1.5155e+12 3729.4

- taxes 1 1.1652e+11 1.5754e+12 3735.7

- list 1 6.5374e+12 7.9963e+12 3998.8

Step: AIC=3724.29

Sale ~ list + bathrooms + taxes + lotsize

Df Sum of Sq RSS AIC

<none> 1.4685e+12 3724.3

- bathrooms 1 2.2847e+10 1.4914e+12 3724.8

- lotsize 1 5.6361e+10 1.5249e+12 3728.4

- taxes 1 1.1050e+11 1.5790e+12 3734.0

- list 1 6.6842e+12 8.1528e+12 4000.0

Call:

lm(formula = Sale ~ list + bathrooms + taxes + lotsize)

Coefficients:

(Intercept) list bathrooms taxes lotsize

7.116e+04 7.609e-01 1.591e+04 1.434e+01 5.750e+00

> step(m1,direction = "backward", k=log(162))

Start: AIC=3763.83

Sale ~ list + bedrooms + bathrooms + lotwidth + lotlength + maxsqfoot +

taxes + location + lotsize

Df Sum of Sq RSS AIC

- location 1 4.3414e+07 1.4566e+12 3758.8

- lotlength 1 4.0049e+08 1.4569e+12 3758.8

- maxsqfoot 1 7.0777e+08 1.4573e+12 3758.8

- lotwidth 1 1.4244e+09 1.4580e+12 3758.9

- bathrooms 1 5.8302e+09 1.4624e+12 3759.4

- bedrooms 1 7.2284e+09 1.4638e+12 3759.5

- lotsize 1 9.9528e+09 1.4665e+12 3759.8

<none> 1.4565e+12 3763.8

- taxes 1 9.1754e+10 1.5483e+12 3768.6

- list 1 6.2570e+12 7.7136e+12 4028.8

Step: AIC=3758.75

Sale ~ list + bedrooms + bathrooms + lotwidth + lotlength + maxsqfoot +

taxes + lotsize

Df Sum of Sq RSS AIC

- lotlength 1 3.7519e+08 1.4570e+12 3753.7

- maxsqfoot 1 6.6588e+08 1.4573e+12 3753.7

- lotwidth 1 1.4106e+09 1.4580e+12 3753.8

- bathrooms 1 5.9660e+09 1.4626e+12 3754.3

- bedrooms 1 7.1972e+09 1.4638e+12 3754.5

- lotsize 1 9.9282e+09 1.4665e+12 3754.8

<none> 1.4566e+12 3758.8

- taxes 1 9.3091e+10 1.5497e+12 3763.7

- list 1 6.2864e+12 7.7430e+12 4024.3

Step: AIC=3753.71

Sale ~ list + bedrooms + bathrooms + lotwidth + maxsqfoot + taxes +

lotsize

Df Sum of Sq RSS AIC

- maxsqfoot 1 6.0072e+08 1.4576e+12 3748.7

- lotwidth 1 1.4436e+09 1.4584e+12 3748.8

- bathrooms 1 6.7364e+09 1.4637e+12 3749.4

- bedrooms 1 6.9494e+09 1.4639e+12 3749.4

- lotsize 1 4.4471e+10 1.5014e+12 3753.5

<none> 1.4570e+12 3753.7

- taxes 1 9.4352e+10 1.5513e+12 3758.8

- list 1 6.2861e+12 7.7431e+12 4019.2

Step: AIC=3748.68

Sale ~ list + bedrooms + bathrooms + lotwidth + taxes + lotsize

Df Sum of Sq RSS AIC

- lotwidth 1 1.3002e+09 1.4589e+12 3743.7

- bedrooms 1 8.4027e+09 1.4660e+12 3744.5

- bathrooms 1 9.3962e+09 1.4670e+12 3744.6

- lotsize 1 4.3912e+10 1.5015e+12 3748.4

<none> 1.4576e+12 3748.7

- taxes 1 1.1766e+11 1.5752e+12 3756.2

- list 1 6.5242e+12 7.9817e+12 4019.1

Step: AIC=3743.74

Sale ~ list + bedrooms + bathrooms + taxes + lotsize

Df Sum of Sq RSS AIC

- bedrooms 1 9.6765e+09 1.4685e+12 3739.7

- bathrooms 1 9.7008e+09 1.4686e+12 3739.7

<none> 1.4589e+12 3743.7

- lotsize 1 5.6582e+10 1.5155e+12 3744.8

- taxes 1 1.1652e+11 1.5754e+12 3751.1

- list 1 6.5374e+12 7.9963e+12 4014.3

Step: AIC=3739.72

Sale ~ list + bathrooms + taxes + lotsize

Df Sum of Sq RSS AIC

- bathrooms 1 2.2847e+10 1.4914e+12 3737.1

<none> 1.4685e+12 3739.7

- lotsize 1 5.6361e+10 1.5249e+12 3740.7

- taxes 1 1.1050e+11 1.5790e+12 3746.4

- list 1 6.6842e+12 8.1528e+12 4012.3

Step: AIC=3737.14

Sale ~ list + taxes + lotsize

Df Sum of Sq RSS AIC

- lotsize 1 3.8225e+10 1.5296e+12 3736.2

<none> 1.4914e+12 3737.1

- taxes 1 1.3054e+11 1.6219e+12 3745.6

- list 1 9.4650e+12 1.0956e+13 4055.1

Step: AIC=3736.15

Sale ~ list + taxes

Df Sum of Sq RSS AIC

<none> 1.5296e+12 3736.2

- taxes 1 1.4758e+11 1.6772e+12 3746.0

- list 1 1.0430e+13 1.1960e+13 4064.2

Call:

lm(formula = Sale ~ list + taxes)

Coefficients:

(Intercept) list taxes

1.167e+05 7.957e-01 1.626e+01