Homework 7

Alexander Van Roijen

February 23, 2019

```
cellData = read.csv('cells.csv')
salesData = read.csv('Sales.csv')
```

Problem 1

```
basecelllm = lm(count1~dose+sex+age+count0,cellData)
print(summary(basecelllm))
```

```
##
## Call:
## lm(formula = count1 ~ dose + sex + age + count0, data = cellData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -563.35  -97.33   12.82   69.38 1278.59
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 274.3173   174.4802   1.572   0.1249
## dose          5.8104     1.0213   5.689 1.98e-06 ***
## sex          26.2302    96.0152   0.273   0.7863
## age          -5.9333     3.1161  -1.904   0.0651 .
## count0        0.8150     0.7534   1.082   0.2868
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 288.9 on 35 degrees of freedom
## Multiple R-squared:  0.5225, Adjusted R-squared:  0.468
## F-statistic: 9.576 on 4 and 35 DF,  p-value: 2.441e-05
```

Clearly, we can see from our model that with adjustment for age, sex, and pre treatment cell count, dose is still significant in the model. Directly, we will say that according to our model, if we hold pretreatment cell count, age, and sex constant, we expect to see dose to make a mean difference in post treatment cell count by 5.8 per unit difference in does.

Now lets test a null hypothesis on $\beta_{dose} = 0$.

```
print(names(summary(basecelllm)))
```

```
##  [1] "call"          "terms"         "residuals"     "coefficients"
##  [5] "aliased"       "sigma"         "df"            "r.squared"
##  [9] "adj.r.squared" "fstatistic"    "cov.unscaled"
```

```
hold = summary(basecelllm)
wiw = coef(hold)[, 2]
doseCoef = coef(hold)[, 1][2]
dosestder=wiw[2]
zval = abs(doseCoef/dosestder)
```

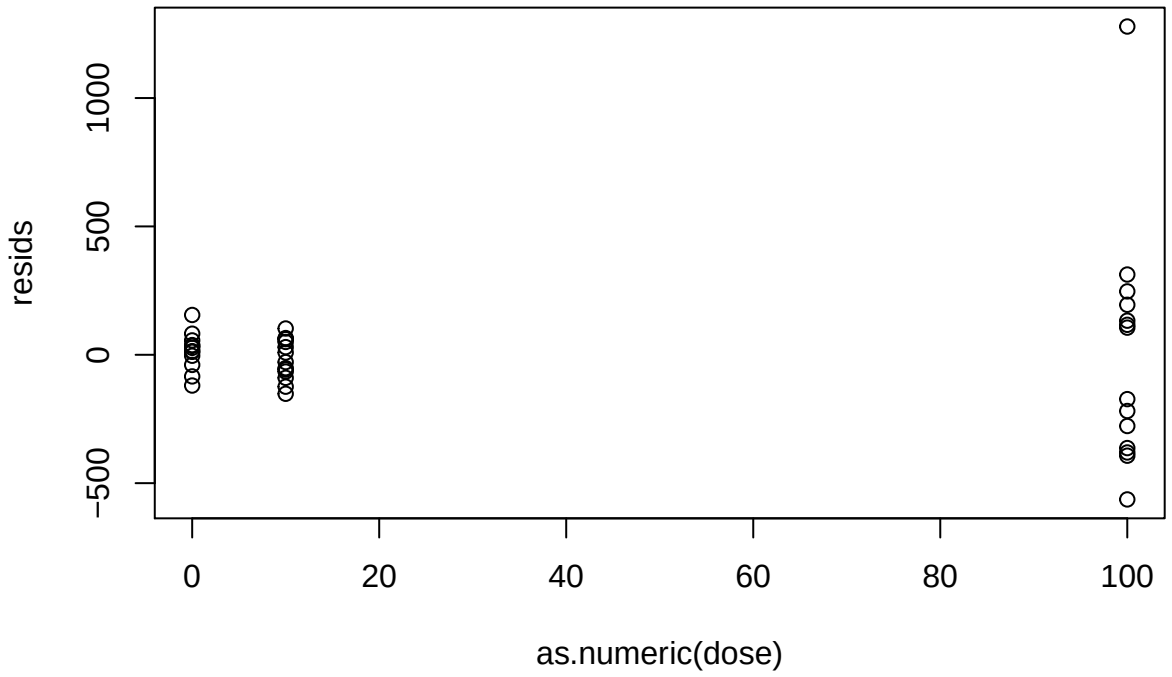
```r
n=nrow(cellData)
p = 2*(1-pt(zval,df=n-5))
print(p)
```

```
##         dose
## 1.982165e-06
```

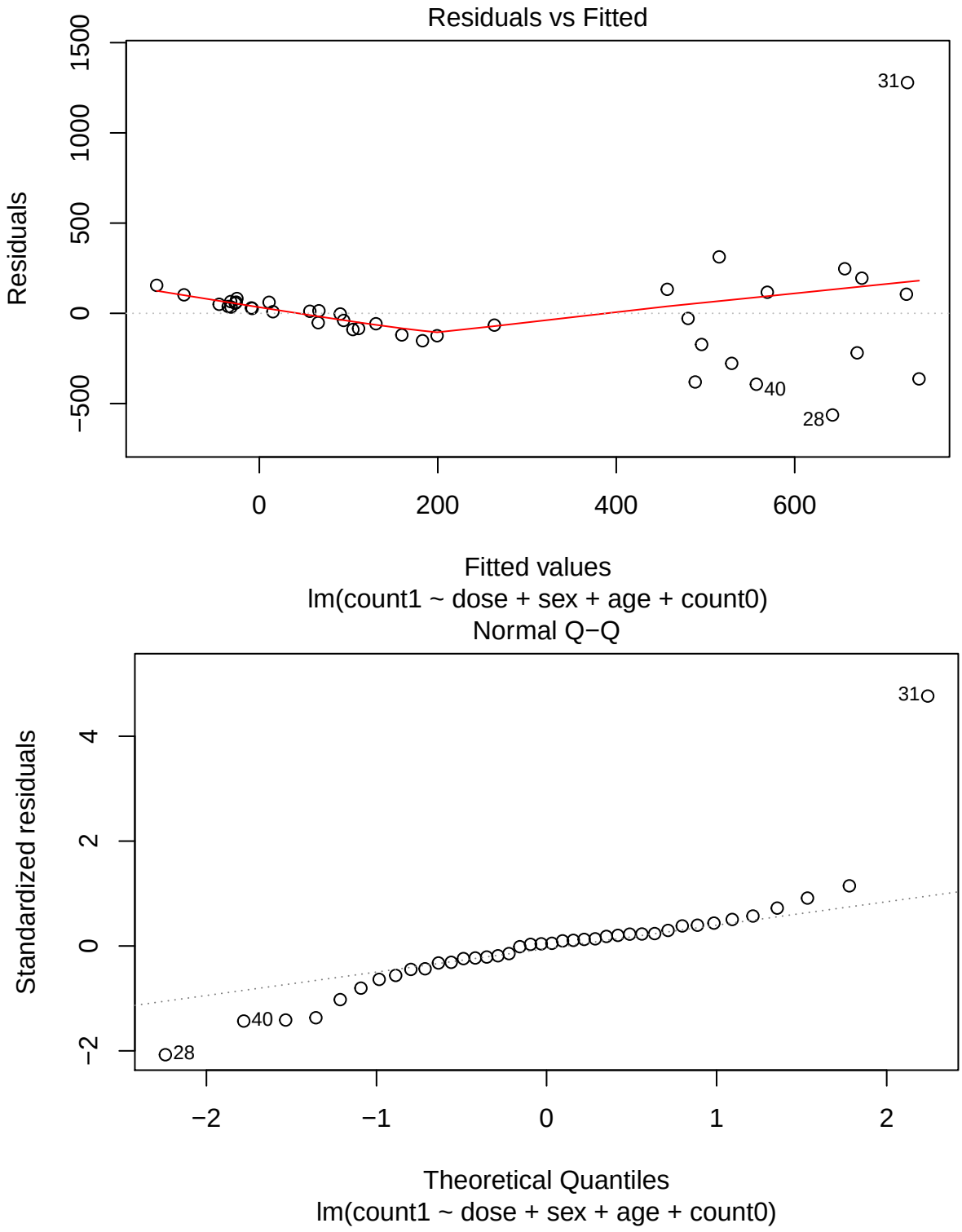
According to this null hypothesis test, we can say with good confidence that we can reject the null hypothesis that dose has no effect on post treatment cell count even with adjustment on sex age and pretreatment cell count.

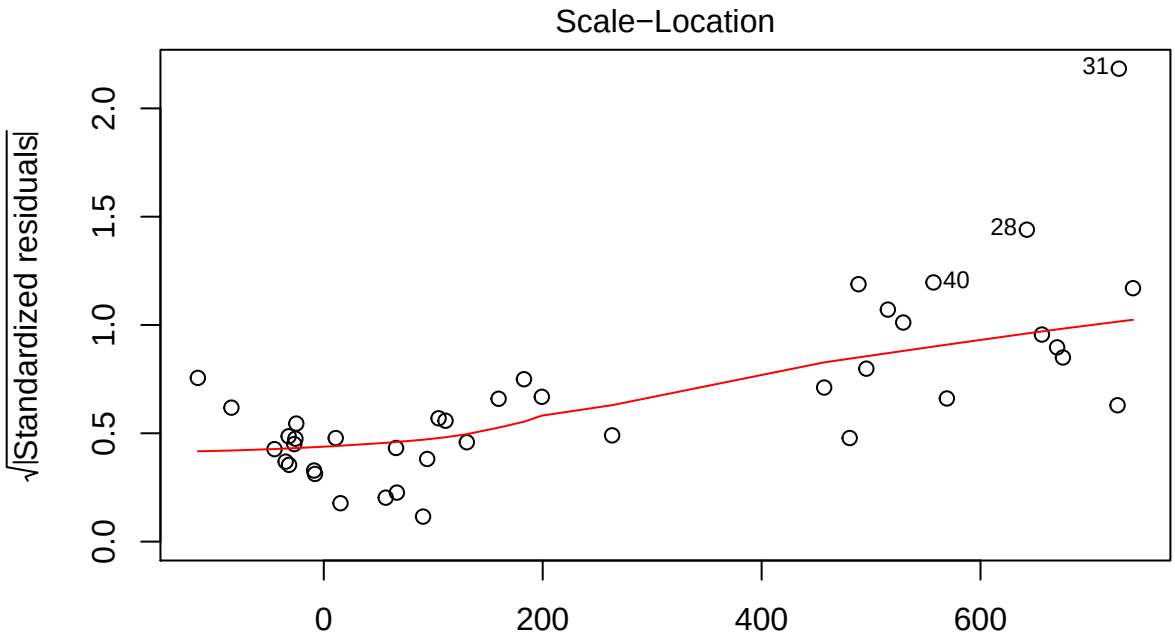
However, now we must consider if our data handles the assumptions of independence, normaliy, linearity, and constant variance. For now, we will consider the independence assumptions handled, but this isnt always garunteed.

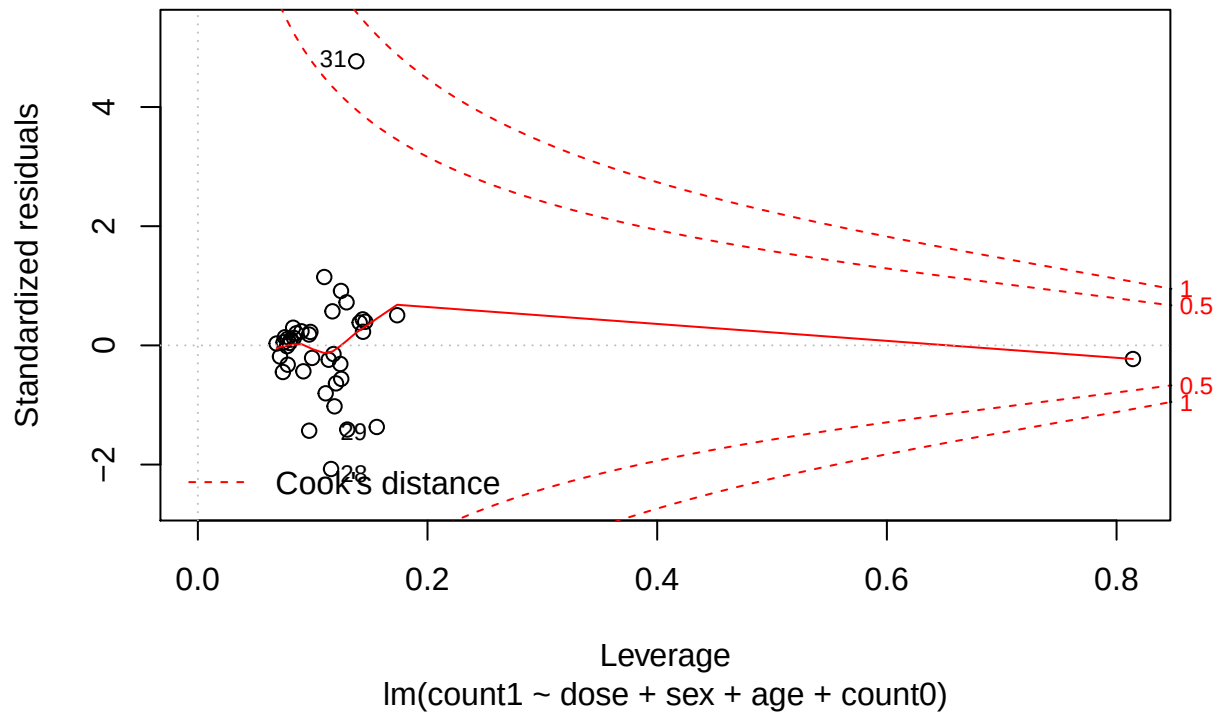
```r
resids = hold$residuals
plot(resids~as.numeric(dose),data=cellData)
```



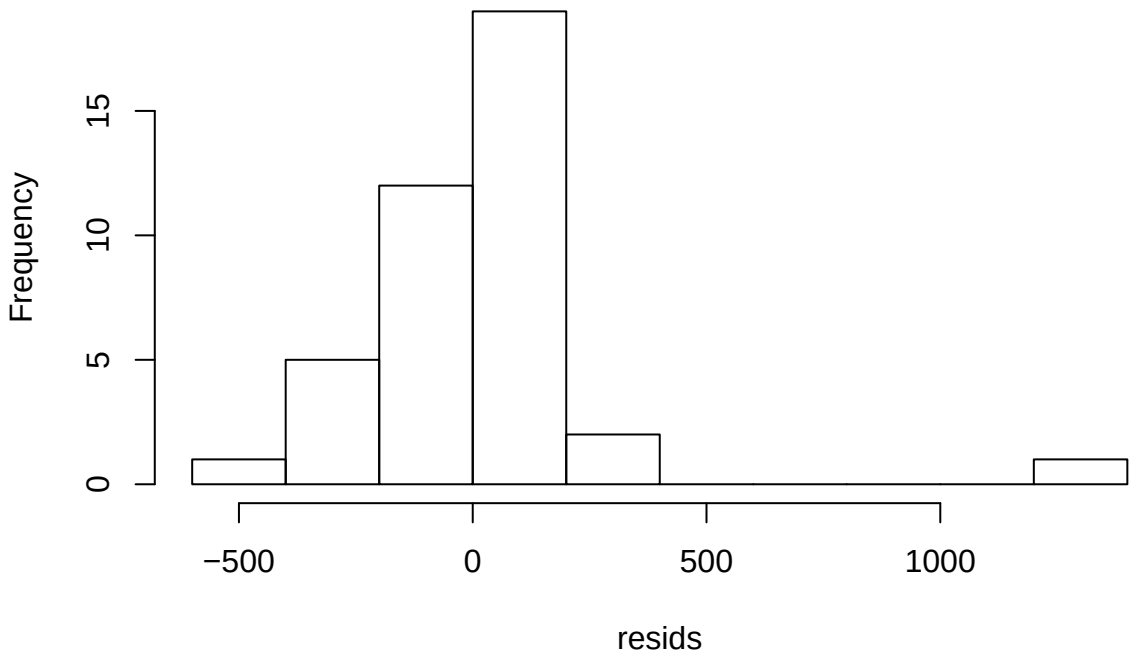

```r
plot(basecelllm)
```


Residuals vs Fitted
Residuals
1500
1000
500
0
−500
31
40
28
0
200
400
600
Fitted values
lm(count1 ~ dose + sex + age + count0)
Normal Q−Q
Standardized residuals
4
2
0
−2
31
40
28
−2
−1
0
1
2
Theoretical Quantiles
lm(count1 ~ dose + sex + age + count0)


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
0
200
400
600
31
28
40
Fitted values
lm(count1 ~ dose + sex + age + count0)


Residuals vs Leverage
Standardized residuals
−2
0
2
4
0.0
0.2
0.4
0.6
0.8
31
29
28
Cook's distance
1
0.5
0.5
1
Leverage
lm(count1 ~ dose + sex + age + count0)

```
hist(resids)
```



From the plots, we have some concern. From the dose versus residual plots, we have some concern as it appears there may be some sign of nonconstant variance. From plotting the fitted values against standardized residuals, we see a positive relationship between mean and variance. Further, from both our qqplot and our residuals histogram, we see a left skew, showing a quite non-normal distribution.

As it stands, perhaps some transformations could be used to solve this issue, but that isnt the objective of this problem. Currently, I would say, most importantly, the lack of constant variance paired with lack of normality makes our inferences from the model invalid.

Problem 2

To begin with, we clearly need to hold some level of adjustment for the other factors of interest when determining the impact of another bathroom. Further, we will first develop an answer at the overall level, and then delve into the subets of houses below and above average size.

```
overallSLM=lm(LAST_SALE_PRICE~BATHS+SQFT+LOT_SIZE+BEDS,salesData)
hold = summary(overallSLM)
print(hold)
```

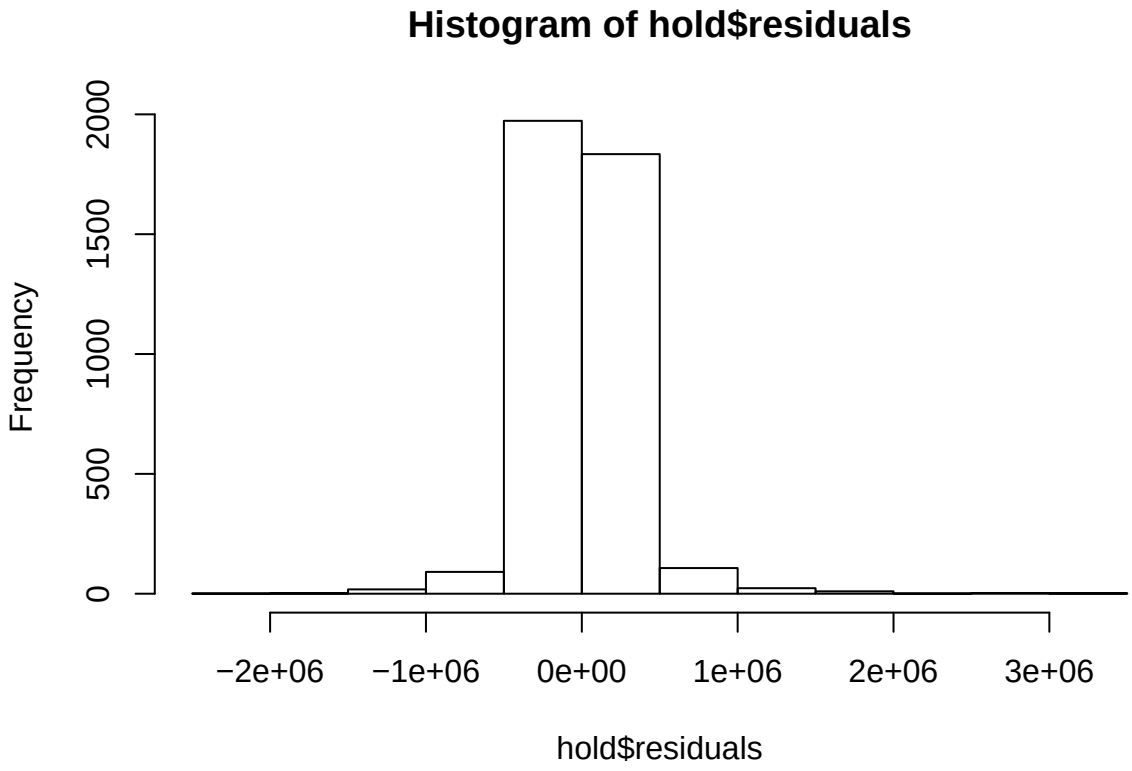
```
##
## Call:
## lm(formula = LAST_SALE_PRICE ~ BATHS + SQFT + LOT_SIZE + BEDS,
##     data = salesData)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2270641  -136780    -6410   113262  3156003
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)  9.364e+04  1.789e+04   5.235 1.74e-07 ***
## BATHS        8.694e+04  8.612e+03  10.095  < 2e-16 ***
## SQFT         3.554e+02  7.841e+00  45.319  < 2e-16 ***
## LOT_SIZE    -2.818e+00  9.736e-01  -2.895  0.00381 **
## BEDS        -9.310e+04  6.665e+03 -13.969  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 300500 on 4060 degrees of freedom
##   (630 observations deleted due to missingness)
## Multiple R-squared:  0.5754, Adjusted R-squared:  0.575
## F-statistic:  1376 on 4 and 4060 DF,  p-value: < 2.2e-16
```

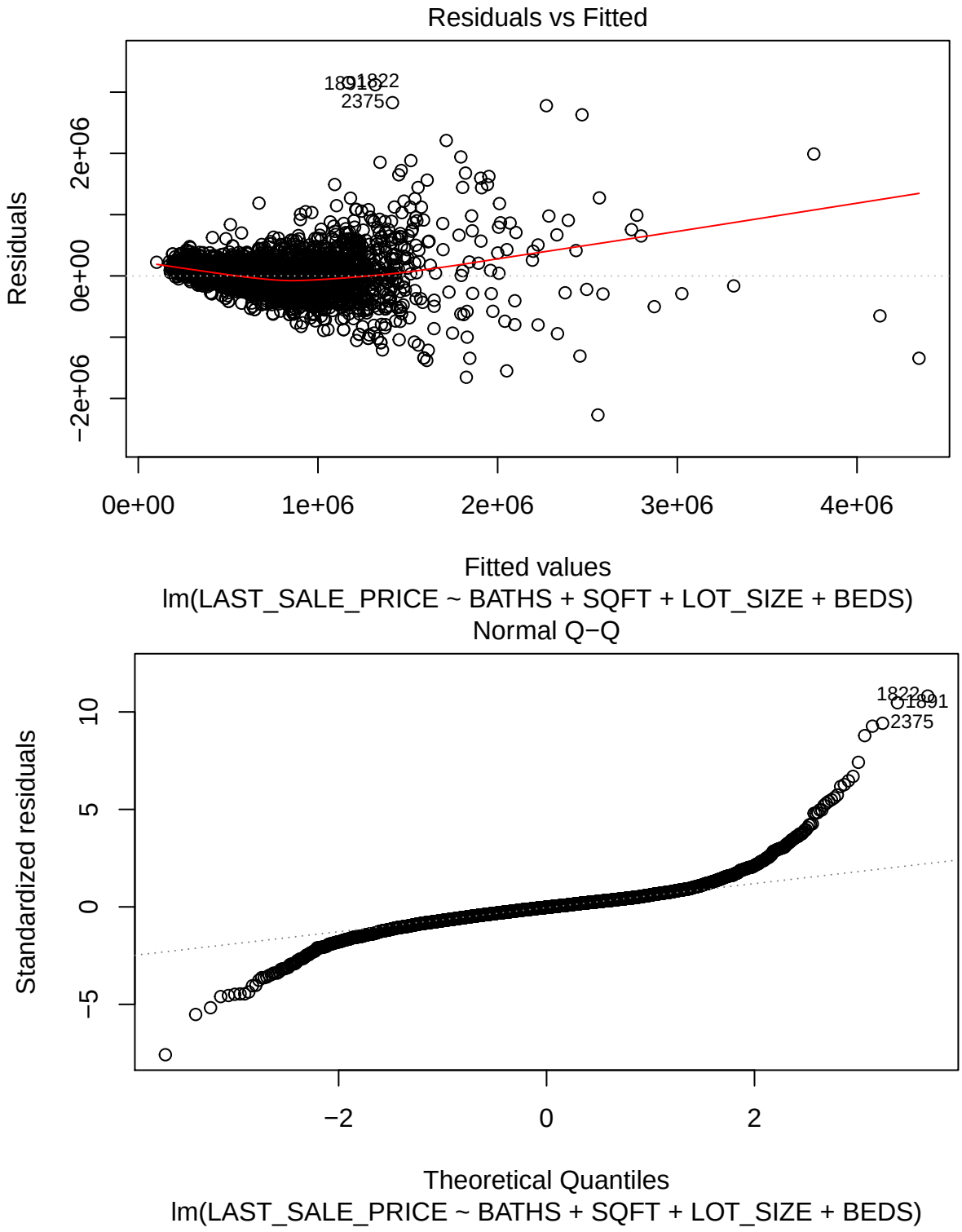
We analyzed this earlier, but again, we see everything is significant in this model. As it stands, we could say that for every additilonal bathroom added to a house, assuming the size, lot_size, and beds remain constant, we would see an increase in mean housevalue of approximatley 87000$.

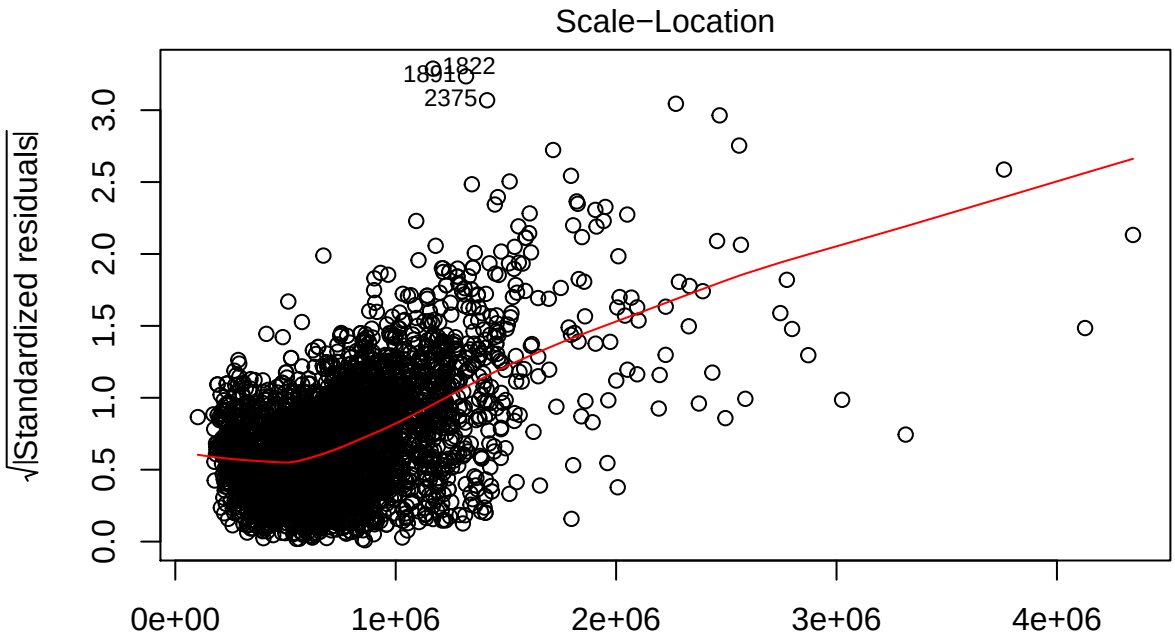
Now lets quickly take a look to see if the assumptions are met.

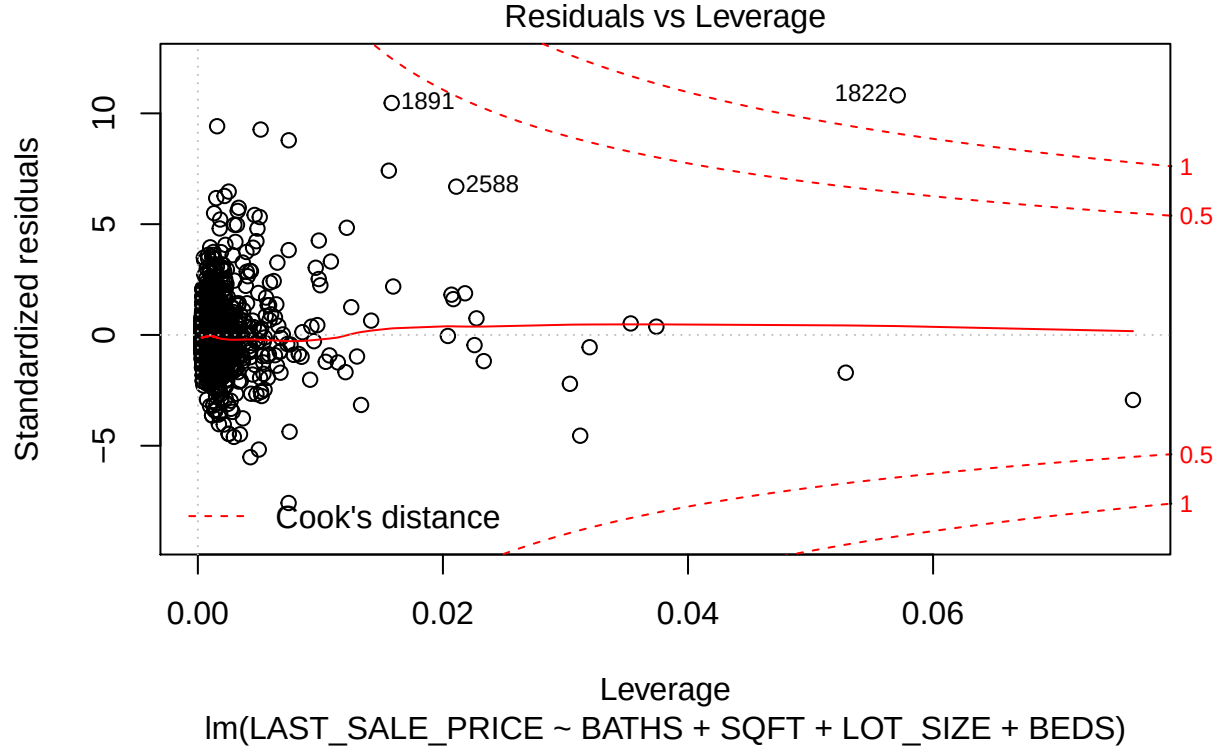
```
hist(hold$residuals)
```




```
plot(overallSLM)
```


Residuals vs Fitted
1891
1822
2375
Residuals
2e+06
0e+00
−2e+06
0e+00
1e+06
2e+06
3e+06
4e+06
Fitted values
lm(LAST_SALE_PRICE ~ BATHS + SQFT + LOT_SIZE + BEDS)
Normal Q−Q
1822
1891
2375
Standardized residuals
10
5
0
−5
−2
0
2
Theoretical Quantiles
lm(LAST_SALE_PRICE ~ BATHS + SQFT + LOT_SIZE + BEDS)






```
print(length(salesData$BATHS))
```

```
## [1] 4695
```

```
print(length(hold$residuals))
```

```
## [1] 4065
```

Clearly,we can see that the normality isnt too well met, and that there seems to be a non constant variance. However, perhaps subsetting the data will solve this issue.

```
meansize = mean(salesData$SQFT,na.rm = TRUE)
#print(sum(is.na(salesData$SQFT)))
#print(meansize)
belowAvg = subset(salesData,SQFT<=meansize)
aboveAvg =  subset(salesData,SQFT>meansize)
#print(length(belowAvg$LAST_SALE_PRICE))
#print(length(aboveAvg$LAST_SALE_PRICE))
#print(length(salesData$LAST_SALE_PRICE))
#these dont add up, should i be concerned?
blm=lm(LAST_SALE_PRICE~BATHS+SQFT+LOT_SIZE+BEDS,belowAvg)
alm=lm(LAST_SALE_PRICE~BATHS+SQFT+LOT_SIZE+BEDS,aboveAvg)
bsum = summary(blm)
asum = summary(alm)
print(bsum)
```

```
##
## Call:
## lm(formula = LAST_SALE_PRICE ~ BATHS + SQFT + LOT_SIZE + BEDS,
##     data = belowAvg)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -523807  -80977  -12984   69613  857853
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.044e+05  1.444e+04  21.079  < 2e-16 ***
## BATHS        4.018e+04  5.727e+03   7.016 3.03e-12 ***
## SQFT         1.887e+02  8.984e+00  21.000  < 2e-16 ***
## LOT_SIZE    -1.260e+01  9.261e-01 -13.605  < 2e-16 ***
## BEDS        -2.203e+04  4.616e+03  -4.772 1.94e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 132500 on 2191 degrees of freedom
##   (468 observations deleted due to missingness)
## Multiple R-squared:  0.3389, Adjusted R-squared:  0.3377
## F-statistic: 280.8 on 4 and 2191 DF,  p-value: < 2.2e-16
```

```
print(asum)
```

```
##
## Call:
## lm(formula = LAST_SALE_PRICE ~ BATHS + SQFT + LOT_SIZE + BEDS,
##     data = aboveAvg)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2527565  -190758   -22423   167812  3110155
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept) -1.161e+05  4.652e+04  -2.495   0.0127 *
## BATHS        1.212e+05  1.563e+04   7.754 1.46e-14 ***
## SQFT         4.068e+02  1.446e+01  28.127  < 2e-16 ***
## LOT_SIZE    -2.684e+00  1.511e+00  -1.776   0.0759 .
## BEDS        -1.077e+05  1.241e+04  -8.681  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 406900 on 1864 degrees of freedom
##   (138 observations deleted due to missingness)
## Multiple R-squared:  0.4872, Adjusted R-squared:  0.4861
## F-statistic: 442.8 on 4 and 1864 DF,  p-value: < 2.2e-16
```

Now when we consider the two models seperately, we see baths are still highly significant with different values for their coefficients. For houses with a below average size, we would expect, with lot size, beds, and sqft held constant, to see an average household value increase of 40,000$ per bath inserted into the house. Meanwhile for above average sized houses, we would expect to see a 120 thousand dollar mean value increase per bathroom added.

However, there is another question to be asked here. Is the house increasing or remaining the same in size during the addition of this new bathroom?

Thus, I would ask the owner if their addition will be used in vacant space already within the house, or require additional space in the house to be constructed. In the case of stagnant house sqft size, my answer would remain the same. However, if there was an increase, I would ask for an estimate and give a combined answer. I would say, for below average houses, the value would increase by, assuming all else held constant, by 121210$ for the bathroom and $406.8\$ * sqft$ where $sqft$ is the approximate increase in size of the house.