AMS 317 HW8

Lucy Lin

2022-10-25

Problem 2a The predictor is significant because the F test is able to test for the condition variable compared to the base model (Y ~ 1) and the p value is lower than 0.05.

```
#2a
data = read.csv("kc_house_data.csv")
data$condition = as.factor(data$condition)
fit <- lm(log(price) ~ condition, data = data)
summary(fit)
```

```
##
## Call:
## lm(formula = log(price) ~ condition, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.83120 -0.36487 -0.02839  0.32544  2.84470
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.49117    0.09552 130.764  < 2e-16 ***
## condition2   0.04694    0.10352   0.453     0.65
## condition3   0.56527    0.09563   5.911 3.45e-09 ***
## condition4   0.52086    0.09578   5.438 5.44e-08 ***
## condition5   0.66715    0.09636   6.923 4.54e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5232 on 21608 degrees of freedom
## Multiple R-squared:  0.01385,    Adjusted R-squared:  0.01367
## F-statistic: 75.86 on 4 and 21608 DF,  p-value: < 2.2e-16
```

Problem 2b condition1: 12.49117 condition2: 12.49117 + 0.04694 = 12.53811 condition3: 12.49117 + 0.56527 = 13.05644 condition4: 12.49117 + 0.52086 = 13.01203 condition5: 12.49117 + 0.66715 = 13.15832

Problem 2c

```
data$condition_sc = data$condition
contrasts(data$condition_sc) = contr.sum
fit1 <- lm(log(price) ~ condition_sc, data = data)
summary(fit1)
```

```
##
## Call:
## lm(formula = log(price) ~ condition_sc, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.83120 -0.36487 -0.02839  0.32544  2.84470
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   12.85122    0.02092 614.191  < 2e-16 ***
## condition_sc1 -0.36005    0.07689  -4.682 2.85e-06 ***
## condition_sc2 -0.31311    0.03732  -8.390  < 2e-16 ***
## condition_sc3  0.20523    0.02120   9.680  < 2e-16 ***
## condition_sc4  0.16082    0.02160   7.444 1.01e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5232 on 21608 degrees of freedom
## Multiple R-squared:  0.01385,    Adjusted R-squared:  0.01367
## F-statistic: 75.86 on 4 and 21608 DF,  p-value: < 2.2e-16
```

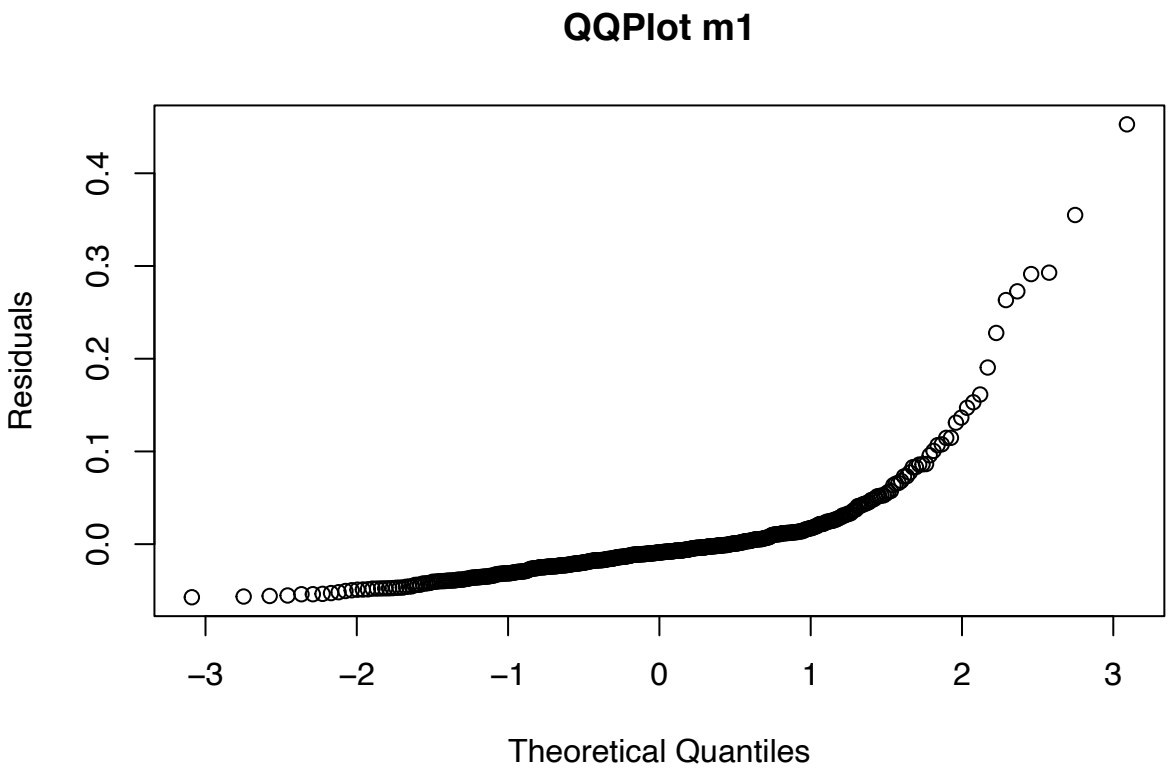
Problem 2d condition1: 12.85122 - 0.36005 = 12.49117 condition2: 12.85122 - 0.31311 = 12.53811 condition3: 12.85122 + 0.20523 = 13.05645 condition4: 12.85122 + 0.16082 = 13.01204 (rounded?) condition5: 12.85122 -(- 0.36005 - 0.31311 + 0.20523 + 0.16082) = 13.15833

The averages are the same as the ones in 2b.

Problem 3a It is very right-skewed.

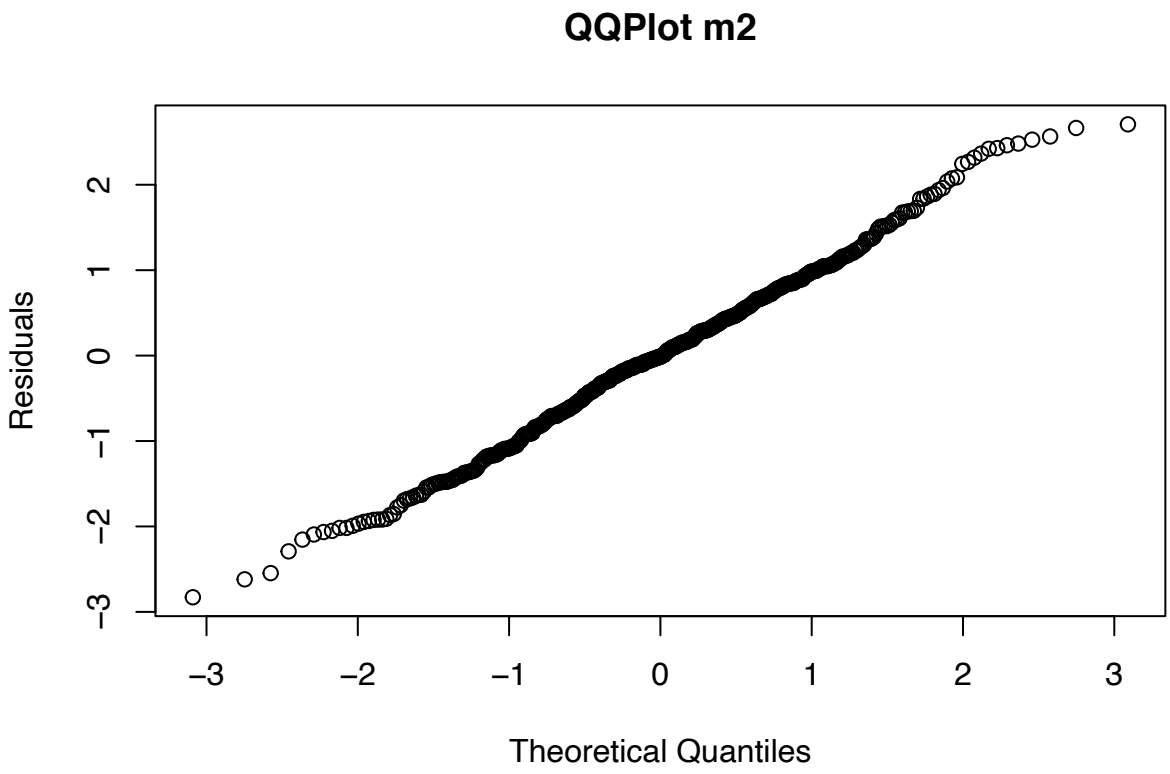
```
set.seed(123)
x1 = runif(500)
x2 = ifelse(runif(500) > 0.5, 1, 0)
y = exp(-5 + x1 + x2 + rnorm(500))

m1 <- lm(y~ x1 + x2)
qqnorm(m1$residuals, ylab = "Residuals", main = "QQPlot m1")
```



Problem 3b This qqplot is more normal than m1.

```
m2 <- lm(log(y)~ x1 + x2)
qqnorm(m2$residuals, ylab = "Residuals", main = "QQPlot m2")
```

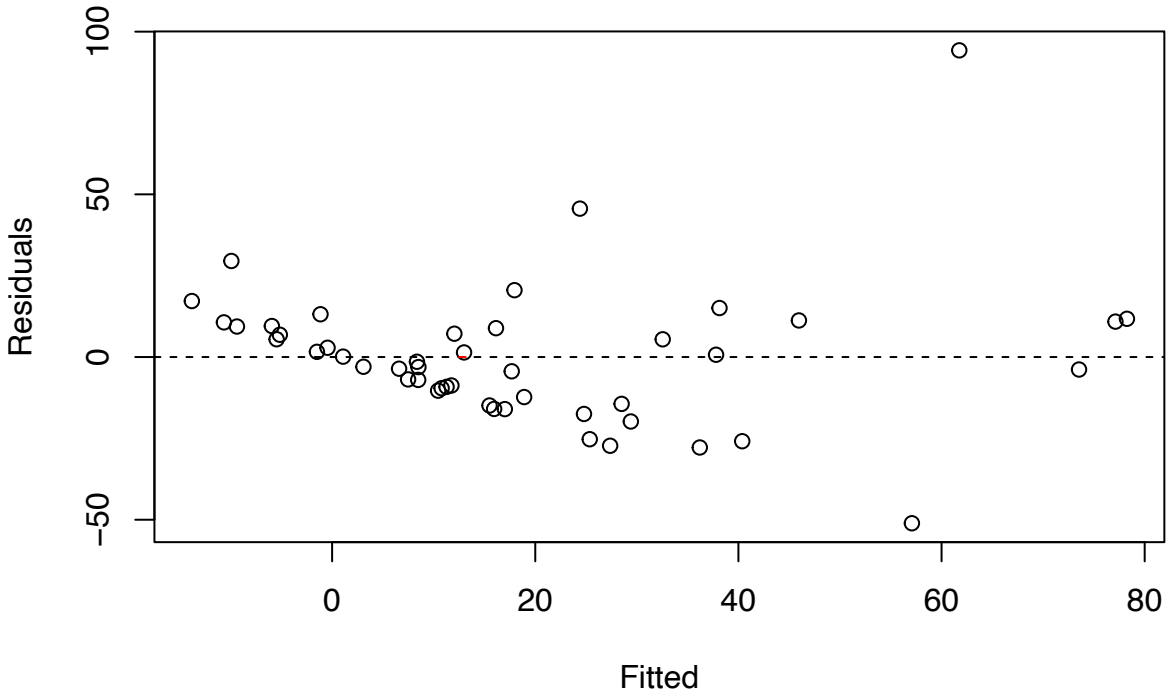


Problem 4 uses the professor's R code from R Note 6. Problem 4a It appears that the variance is not constant.

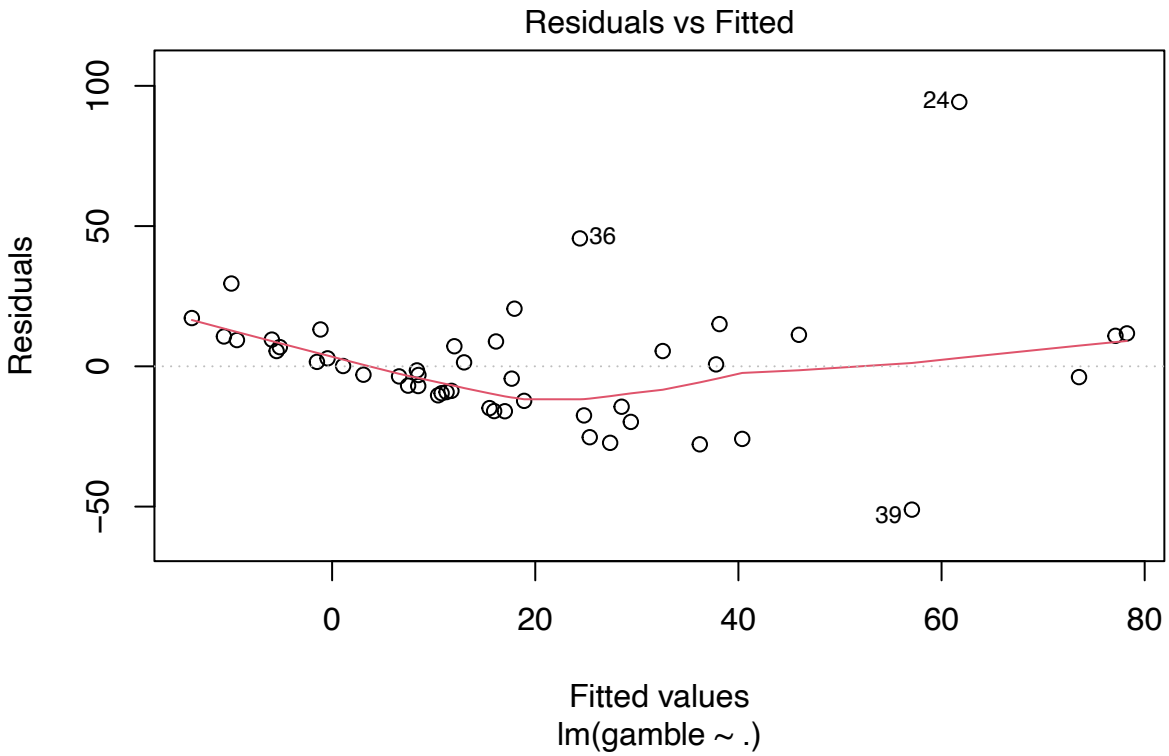
```
data(teengamb, package = 'faraway')
head(teengamb)
```

```
##   sex status income verbal gamble
## 1   1     51   2.00      8    0.0
## 2   1     28   2.50      8    0.0
## 3   1     37   2.00      6    0.0
## 4   1     28   7.00      4    7.3
## 5   1     65   2.00      8   19.6
## 6   1     61   3.47      6    0.1
```

```
fit2 <- lm(gamble ~ ., data = teengamb)
plot(fit2$fitted.values, fit2$residuals, xlab = "Fitted", ylab = "Residuals")
abline(h = 0, lty = 2); lines(lowess(fitted.values(fit), residuals(fit)), col = "red")
```

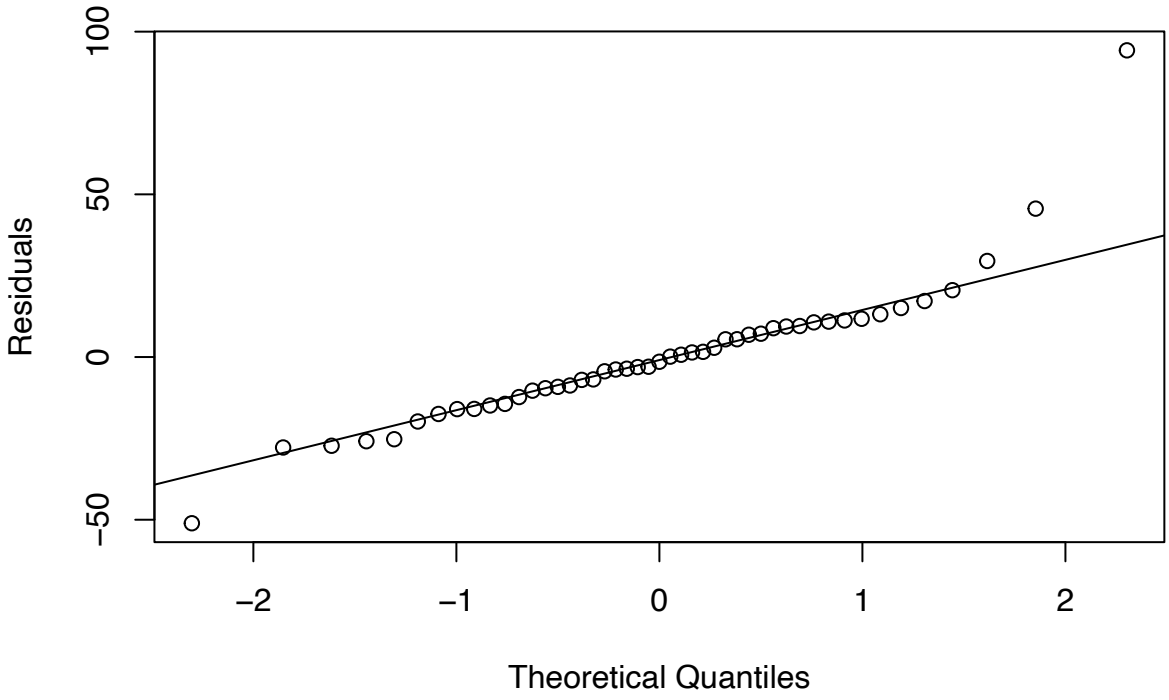



```
plot(fit2, which = 1) # build-in plots have this figure, see ?plot.lm
```

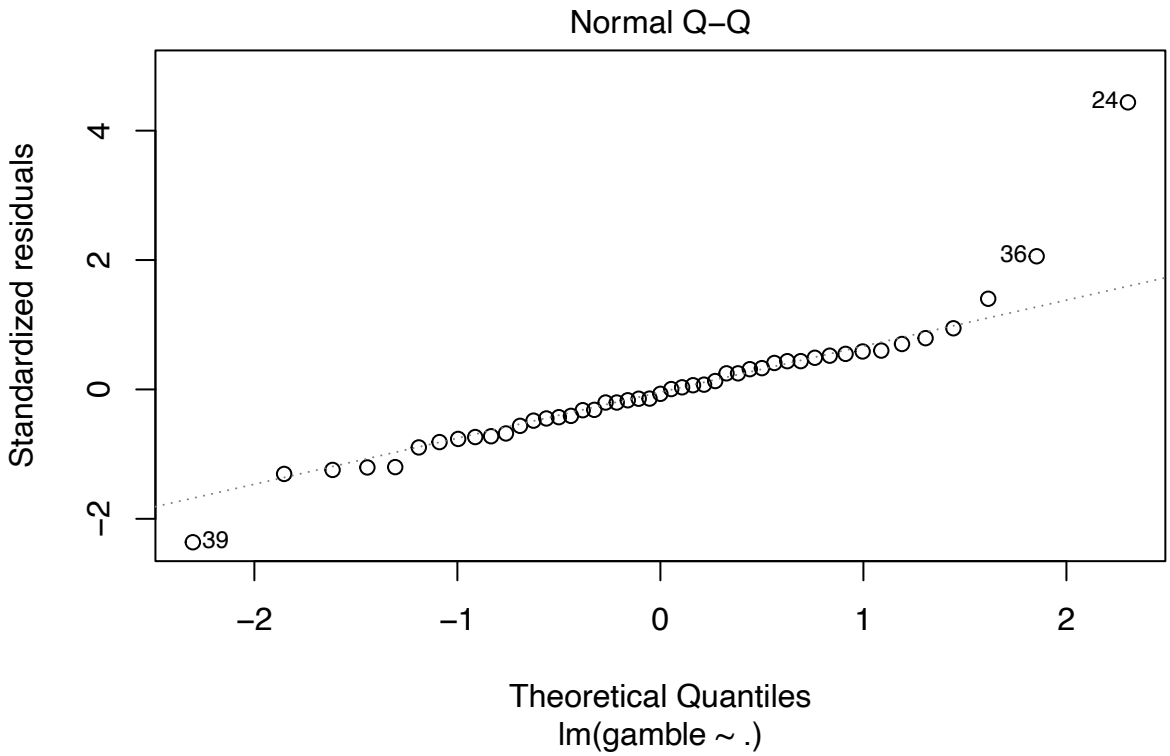


Problem 4b The errors lie mostly along the diagonal line so they are normal.

```
qqnorm(fit2$residuals, ylab = "Residuals", main = "")
qqline(fit2$residuals)
```




```
plot(fit2, which = 2)
```



Problem 4c Point 24 is an outlier because $6.016116 > 3.522795$. There is only one outlier.

```
stud = rstudent(fit2)
stud[which.max(abs(stud))]
```

```
##       24
## 6.016116
```

```
n = dim(teengamb)[1]
p = length(fit2$coefficients)
abs(qt((0.05/n)/2, n-p-1))
```

```
## [1] 3.522795
```

```
stud[which.max(stud > abs(qt((0.05/n)/2, n-p-1)) )]
```

```
##       24
## 6.016116
```

Problem 4d

```
h_diag = hatvalues(fit2)
sum(h_diag)
```

```
## [1] 5
```

```
#p = 5
h_diag[which(h_diag > (2*p/n))]
```

```
##        31        33        35        42
## 0.2395031 0.2213439 0.3118029 0.3016088
```

```
#4 leverage points
```

Problem 4e

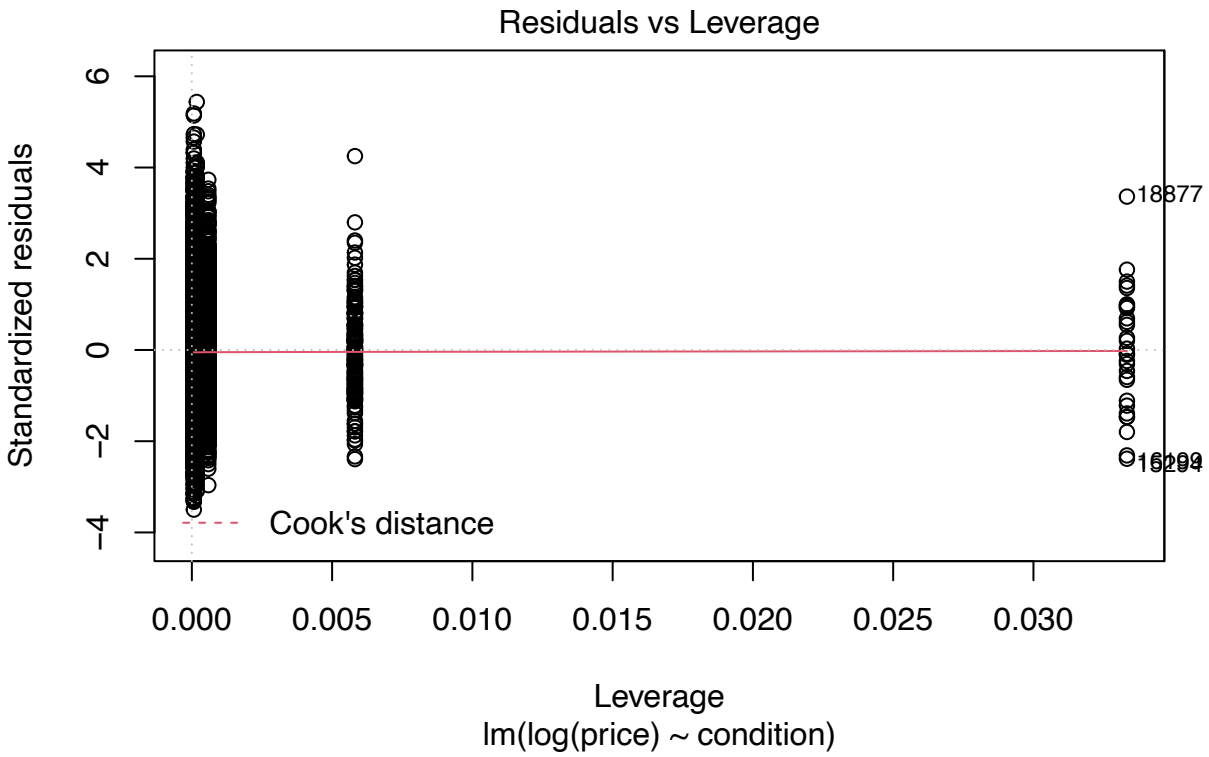
```
stand = rstandard(fit2)
cook = cooks.distance(fit2)
# calculate by ourselves
D = (stand^2/p) * (h_diag/(1-h_diag))
cook - D
```

```
##             1             2             3             4             5
## -8.673617e-19 -1.734723e-18  3.252607e-19  3.469447e-18 -1.387779e-17
##             6             7             8             9            10
##  1.084202e-19  4.336809e-19  8.673617e-19  6.505213e-19 -8.673617e-19
##            11            12            13            14            15
##  0.000000e+00  5.421011e-20 -4.235165e-22 -8.673617e-19 -8.131516e-20
##            16            17            18            19            20
## -5.204170e-18 -1.387779e-17 -2.081668e-17 -8.673617e-19  0.000000e+00
##            21            22            23            24            25
##  0.000000e+00 -1.734723e-18  6.938894e-18 -1.110223e-16  3.388132e-21
##            26            27            28            29            30
##  8.673617e-19  6.938894e-18  0.000000e+00  4.336809e-19  0.000000e+00
##            31            32            33            34            35
##  0.000000e+00 -1.734723e-18 -3.469447e-18  1.084202e-19 -2.081668e-17
##            36            37            38            39            40
## -6.938894e-18  0.000000e+00  0.000000e+00  0.000000e+00 -1.734723e-18
##            41            42            43            44            45
##  0.000000e+00  4.336809e-19 -2.168404e-19  0.000000e+00 -3.252607e-19
##            46            47
## -4.065758e-20 -2.168404e-19
```

```
cook[which(cook > 1)]
```

```
## named numeric(0)
```

```
# no influential points
plot(fit, which = 5)
```


Residuals vs Leverage
Standardized residuals
6
4
2
0
−2
−4
18877
Cook's distance
0.000
0.005
0.010
0.015
0.020
0.025
0.030
Leverage
lm(log(price) ~ condition)