Linear Model Assumptions

Serena Peterson

2023-12-12

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(stats)
library(lmtest)
```

```
## Warning: package 'lmtest' was built under R version 4.3.2
```

```
## Loading required package: zoo
```

```
## Warning: package 'zoo' was built under R version 4.3.2
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

Check Assumptions of Linear Models

```
# Load data
alldata1 <- read.csv(file = file.choose())
alldata2 <- read.csv(file = file.choose())
```

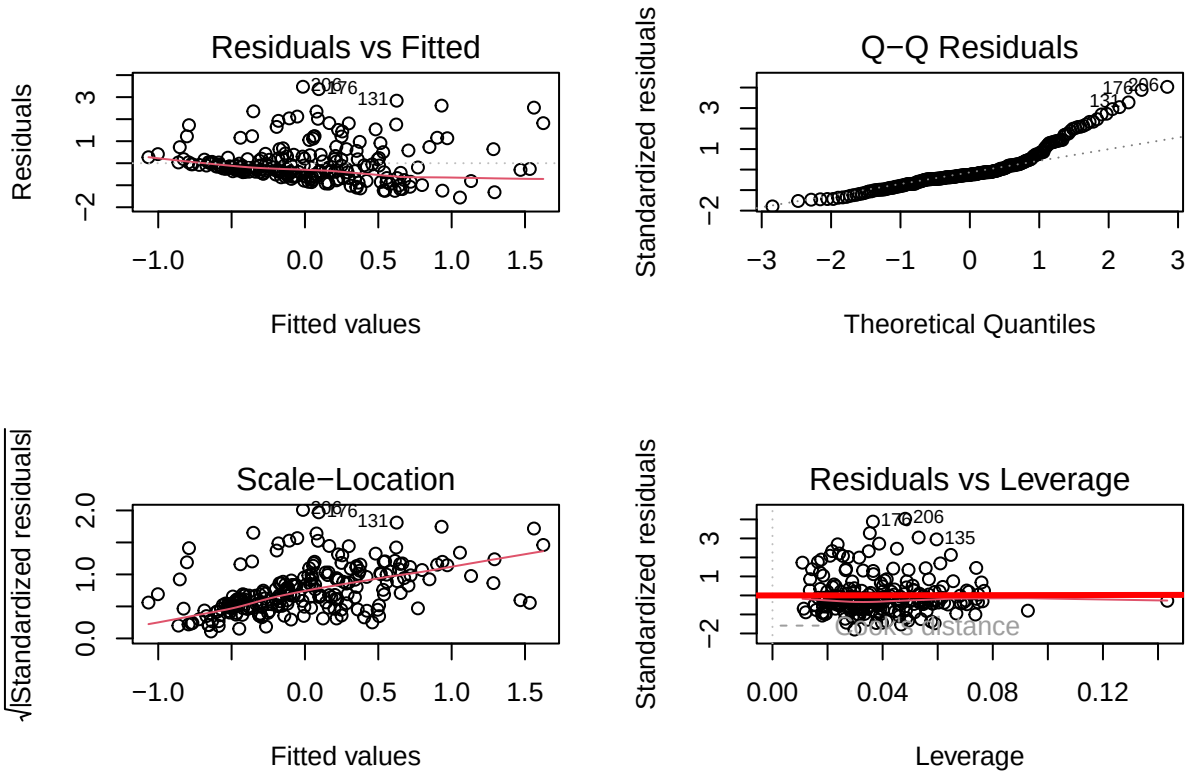
```
#Remove indexing columns
alldata1 <- alldata1 %>% select(-c(X, X.1))
alldata2 <- alldata2 %>% select(-c(X, X.1))
```

```
#Standardize data
alldata1 <-  alldata1 %>% mutate_all(~(scale(.) %>% as.vector))
alldata2 <- alldata2 %>% mutate_all(~(scale(.) %>% as.vector))
```

Data from Approach 1

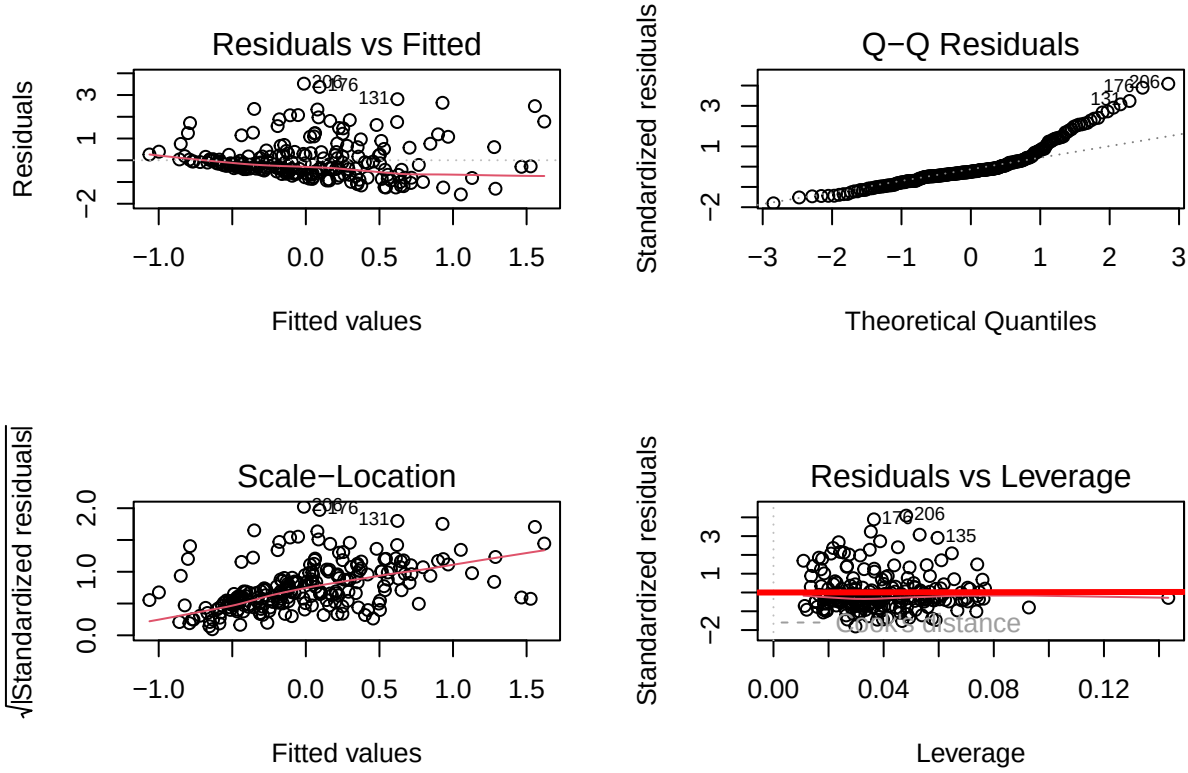
```
#Build models
model1 <- lm(Avg.Loss ~ .,
             data = alldata1 %>% select(-c(RMSE,MAE)))
model2 <- lm(MAE ~ .,
             data = alldata1 %>% select(-c(Avg.Loss,RMSE)))
model3 <- lm(RMSE ~ .,
             data = alldata1 %>% select(-c(Avg.Loss,MAE)))

#Diagnostic Plots
par(mfrow=c(2,2))
plot(model1)
abline(model1, lwd=3, col='red')
```

```
## Warning in abline(model1, lwd = 3, col = "red"): only using the first two of 9
## regression coefficients
```



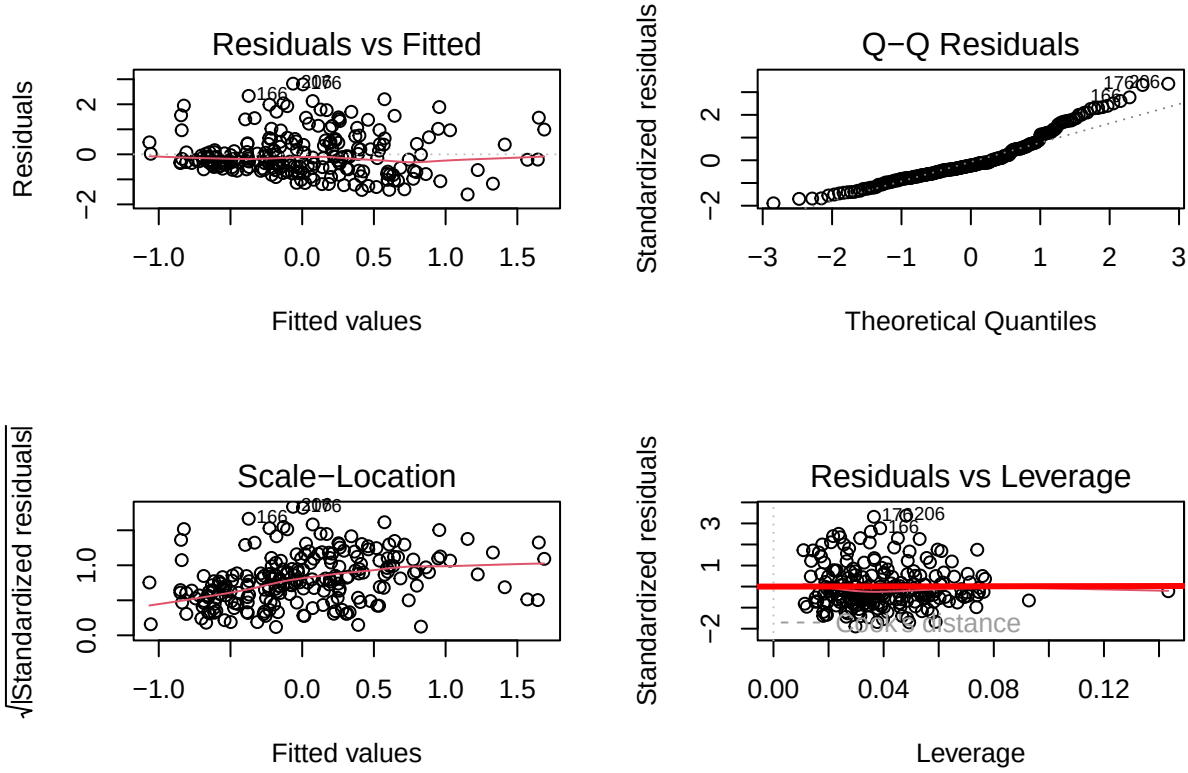

```
par(mfrow=c(2,2))
plot(model2)
abline(model2, lwd=3, col='red')
```

```
## Warning in abline(model2, lwd = 3, col = "red"): only using the first two of 9
## regression coefficients
```




```
par(mfrow=c(2,2))
plot(model3)
abline(model3, lwd=3, col='red')
```

```
## Warning in abline(model3, lwd = 3, col = "red"): only using the first two of 9
## regression coefficients
```




```
#Normality
shapiro.test(model1$res)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  model1$res
## W = 0.87021, p-value = 6.385e-13
```

```
shapiro.test(model2$res)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  model2$res
## W = 0.87088, p-value = 6.964e-13
```

```
shapiro.test(model3$res)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  model3$res
## W = 0.93141, p-value = 9.522e-09
```

```
# Equal Variance
bptest(model1)
```

```
##
```

```
##  studentized Breusch-Pagan test
##
## data:  model1
## BP = 26.189, df = 8, p-value = 0.0009748
```

```
bptest(model2)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model2
## BP = 25.887, df = 8, p-value = 0.001098
```

```
bptest(model3)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model3
## BP = 18.844, df = 8, p-value = 0.01572
```

```
#Independence
dwtest(model1)
```

```
##
##  Durbin-Watson test
##
## data:  model1
## DW = 1.3753, p-value = 8.354e-07
## alternative hypothesis: true autocorrelation is greater than 0
```

```
dwtest(model2)
```

```
##
##  Durbin-Watson test
##
## data:  model2
## DW = 1.3722, p-value = 7.443e-07
## alternative hypothesis: true autocorrelation is greater than 0
```

```
dwtest(model3)
```

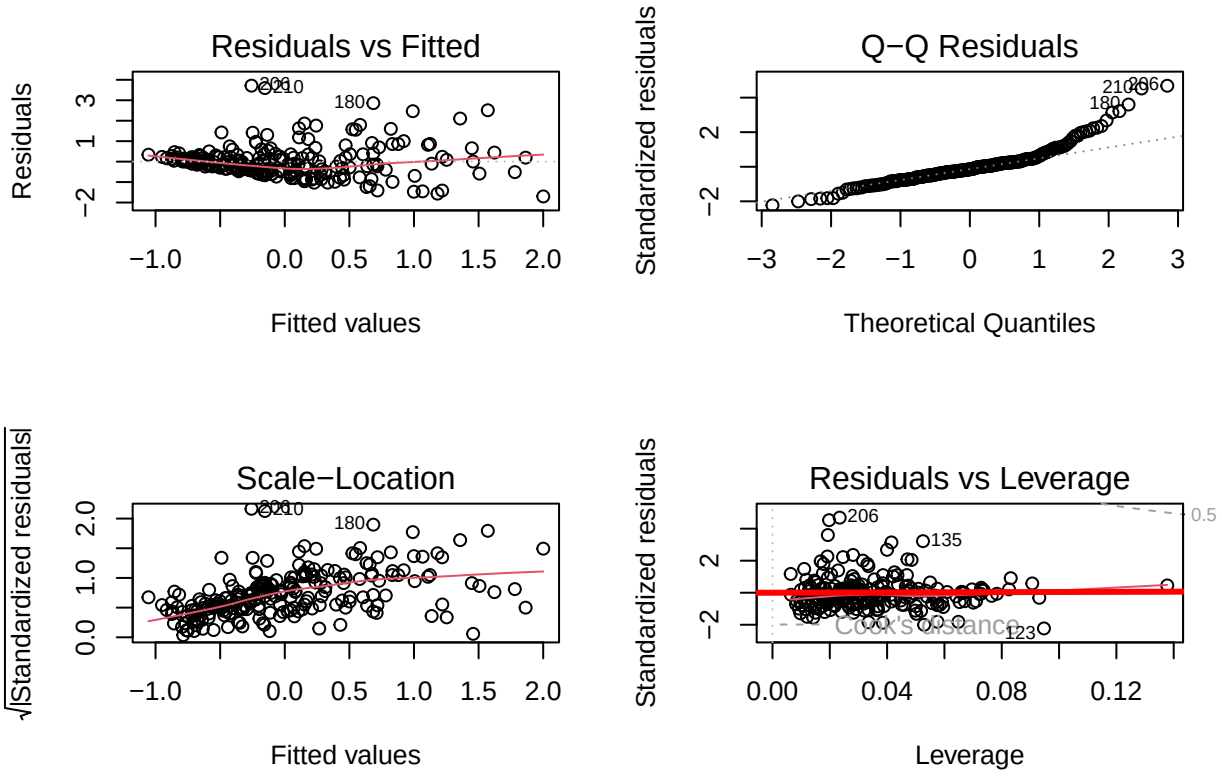
```
##
##  Durbin-Watson test
##
## data:  model3
## DW = 1.226, p-value = 1.634e-09
## alternative hypothesis: true autocorrelation is greater than 0
```

Data from Approach 2

```
#Build Models
model1 <- lm(Avg.Loss ~ .,
             data = alldata2 %>% select(-c(RMSE,MAE)))
model2 <- lm(MAE ~ .,
             data = alldata2 %>% select(-c(Avg.Loss,RMSE)))
model3 <- lm(RMSE ~ .,
             data = alldata2 %>% select(-c(Avg.Loss,MAE)))
```

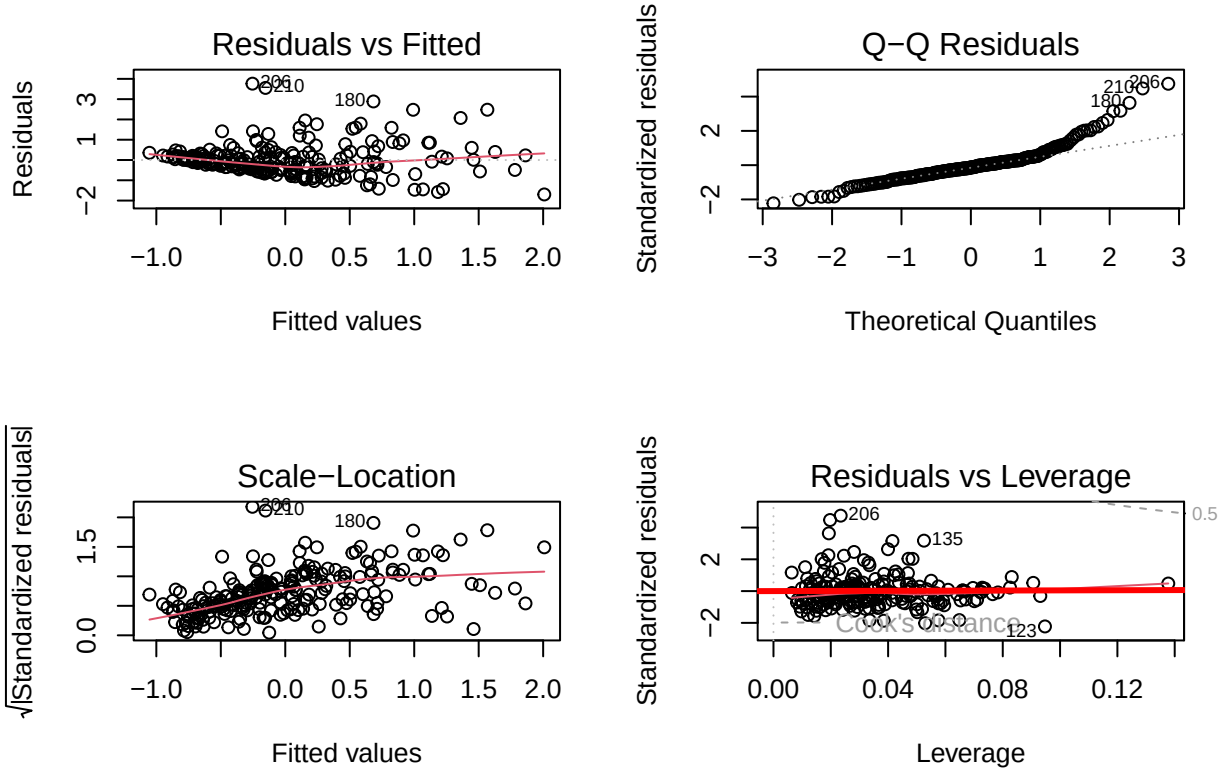
```
#Observations
par(mfrow=c(2,2))
plot(model1)
abline(model1, lwd=3, col='red')
```

```
## Warning in abline(model1, lwd = 3, col = "red"): only using the first two of 8
## regression coefficients
```



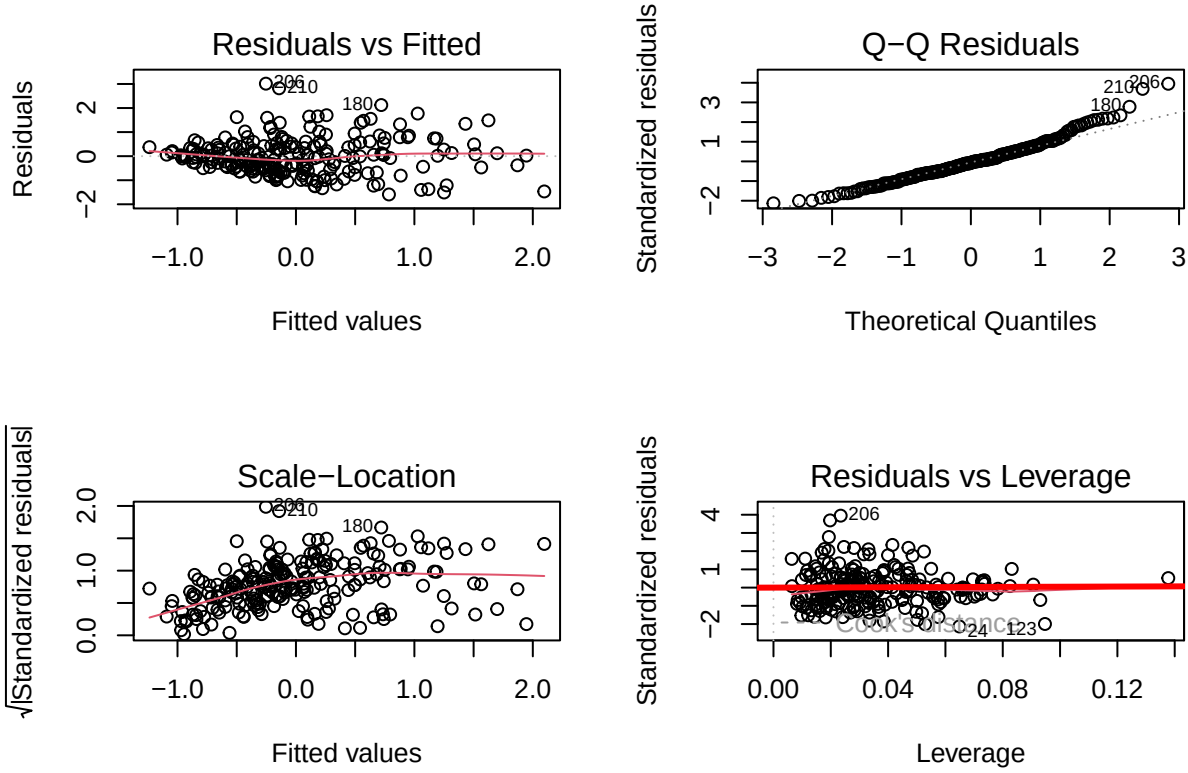

```
par(mfrow=c(2,2))
plot(model2)
abline(model2, lwd=3, col='red')
```

```
## Warning in abline(model2, lwd = 3, col = "red"): only using the first two of 8
## regression coefficients
```




```
par(mfrow=c(2,2))
plot(model3)
abline(model3, lwd=3, col='red')
```

```
## Warning in abline(model3, lwd = 3, col = "red"): only using the first two of 8
## regression coefficients
```




```
#Normality

shapiro.test(model1$res)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  model1$res
## W = 0.88814, p-value = 7.257e-12
```

```
shapiro.test(model2$res)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  model2$res
## W = 0.88797, p-value = 7.086e-12
```

```
shapiro.test(model3$res)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  model3$res
## W = 0.96842, p-value = 6.535e-05
```

```
# Equal Variance
bptest(model1)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model1
## BP = 17.894, df = 7, p-value = 0.01246
```

```
bptest(model2)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model2
## BP = 17.568, df = 7, p-value = 0.01408
```

```
bptest(model3)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model3
## BP = 18.029, df = 7, p-value = 0.01184
```

```
#Independence
dwtest(model1)
```

```
##
##  Durbin-Watson test
##
## data:  model1
## DW = 1.6799, p-value = 0.006268
## alternative hypothesis: true autocorrelation is greater than 0
```

```
dwtest(model2)
```

```
##
##  Durbin-Watson test
##
## data:  model2
## DW = 1.6746, p-value = 0.005596
## alternative hypothesis: true autocorrelation is greater than 0
```

```
dwtest(model3)
```

```
##
##  Durbin-Watson test
##
## data:  model3
## DW = 1.4863, p-value = 3.797e-05
## alternative hypothesis: true autocorrelation is greater than 0
```