STAT625_Project

Brian Zhang, Vi Mai, Xinyu Zhou (Anna), Ziyan Zhao

2023-11-14

```
library(lmerTest)
```

```
## Loading required package: lme4
```

```
## Loading required package: Matrix
```

```
##
## Attaching package: 'lmerTest'
```

```
## The following object is masked from 'package:lme4':
##
##     lmer
```

```
## The following object is masked from 'package:stats':
##
##     step
```

```
library(car)
```

```
## Loading required package: carData
```

```
library(ggplot2)
rm(list = ls())
walmart <- read.csv("Walmart.csv")
head(walmart)
```

```
##   Store       Date Weekly_Sales Holiday_Flag Temperature Fuel_Price      CPI
## 1     1 05-02-2010      1643691            0       42.31      2.572 211.0964
## 2     1 12-02-2010      1641957            1       38.51      2.548 211.2422
## 3     1 19-02-2010      1611968            0       39.93      2.514 211.2891
## 4     1 26-02-2010      1409728            0       46.63      2.561 211.3196
## 5     1 05-03-2010      1554807            0       46.50      2.625 211.3501
## 6     1 12-03-2010      1439542            0       57.79      2.667 211.3806
##   Unemployment
## 1        8.106
## 2        8.106
## 3        8.106
## 4        8.106
## 5        8.106
## 6        8.106
```

Data Preprocessing

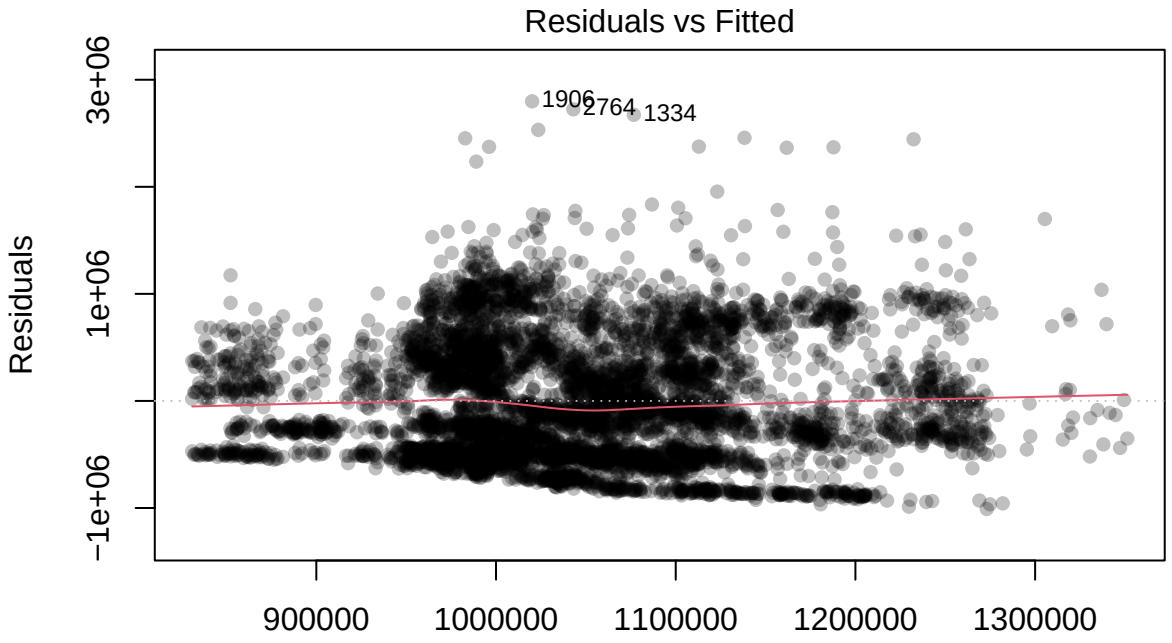
Since dates are strings, they must be converted to parsed and converted to days. Use days since the first day rather than the actual date to make computation easier.

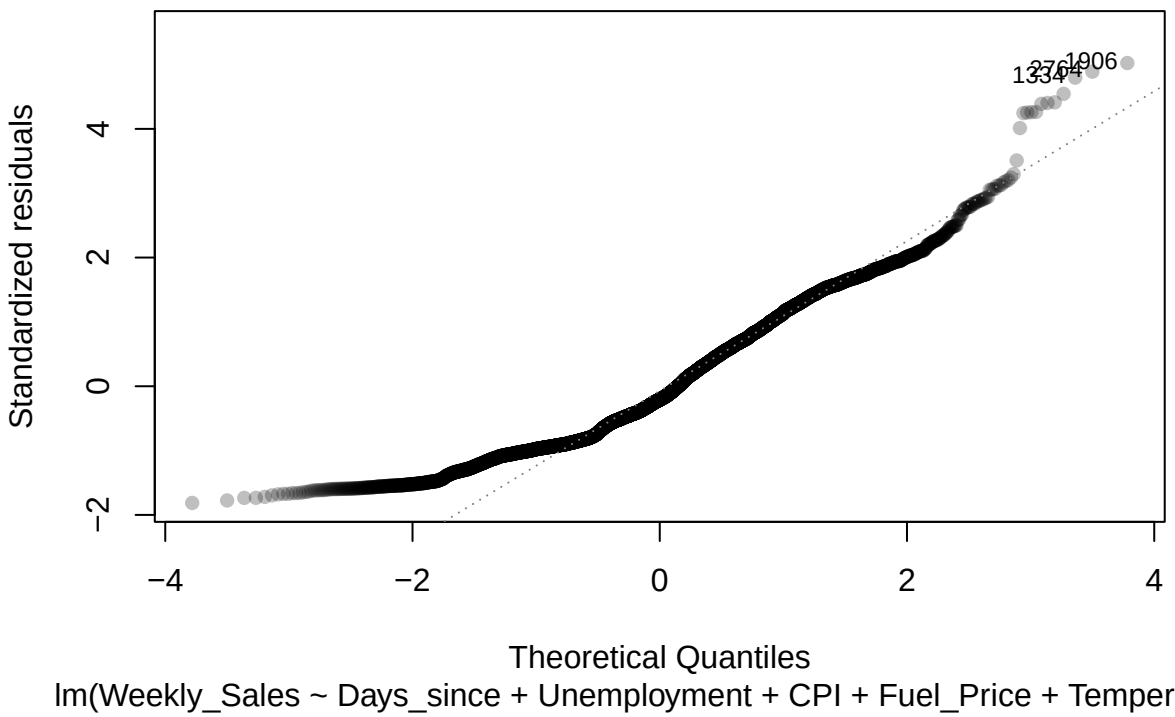
```
# Convert the dates from character strings into days since the first date
asDate_result <- as.Date(walmart$Date, "%d-%m-%Y")
first_date <- min(asDate_result)
days_elapsed <- asDate_result-first_date
walmart["Days_since"] <- days_elapsed
head(walmart)
```

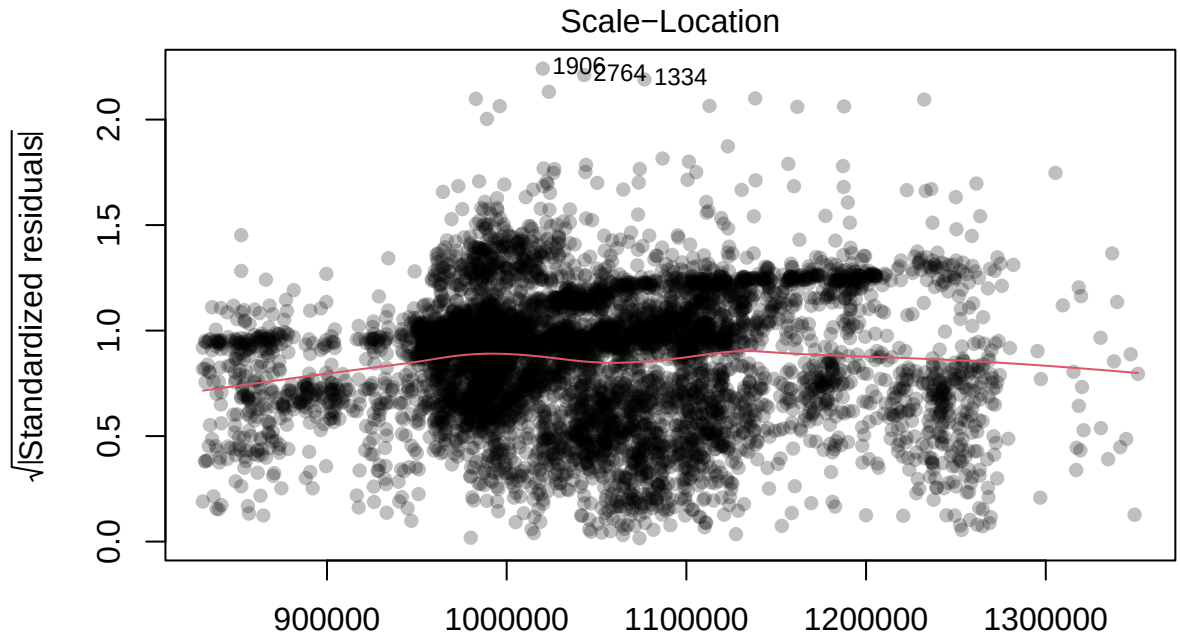
```
##   Store       Date Weekly_Sales Holiday_Flag Temperature Fuel_Price      CPI
## 1     1 05-02-2010      1643691            0       42.31      2.572 211.0964
## 2     1 12-02-2010      1641957            1       38.51      2.548 211.2422
## 3     1 19-02-2010      1611968            0       39.93      2.514 211.2891
## 4     1 26-02-2010      1409728            0       46.63      2.561 211.3196
## 5     1 05-03-2010      1554807            0       46.50      2.625 211.3501
## 6     1 12-03-2010      1439542            0       57.79      2.667 211.3806
##   Unemployment Days_since
## 1        8.106     0 days
## 2        8.106     7 days
## 3        8.106    14 days
## 4        8.106    21 days
## 5        8.106    28 days
## 6        8.106    35 days
```

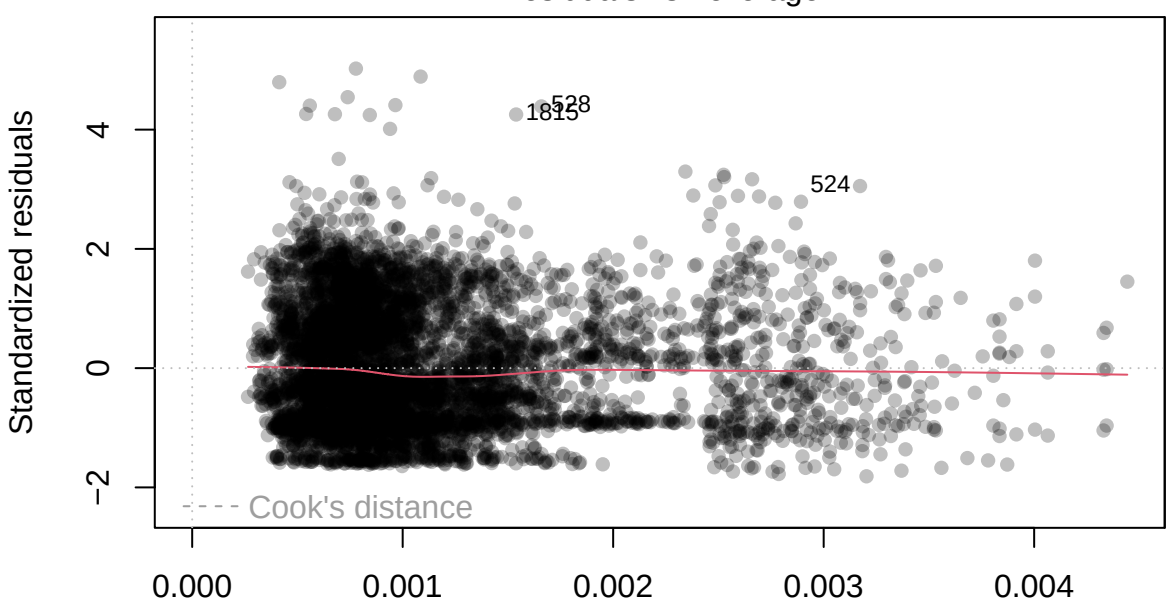
Initial Diagnostic Plots

```
# Comparing the need for a log transformation on Weekly Sales
full_mod_no_log <- lm(Weekly_Sales ~ Days_since + Unemployment + CPI + Fuel_Price + Temperature + Holida
plot(full_mod_no_log,
     col = rgb(red = 0, green = 0, blue = 0, alpha = 0.25),
     pch = 16
)
```

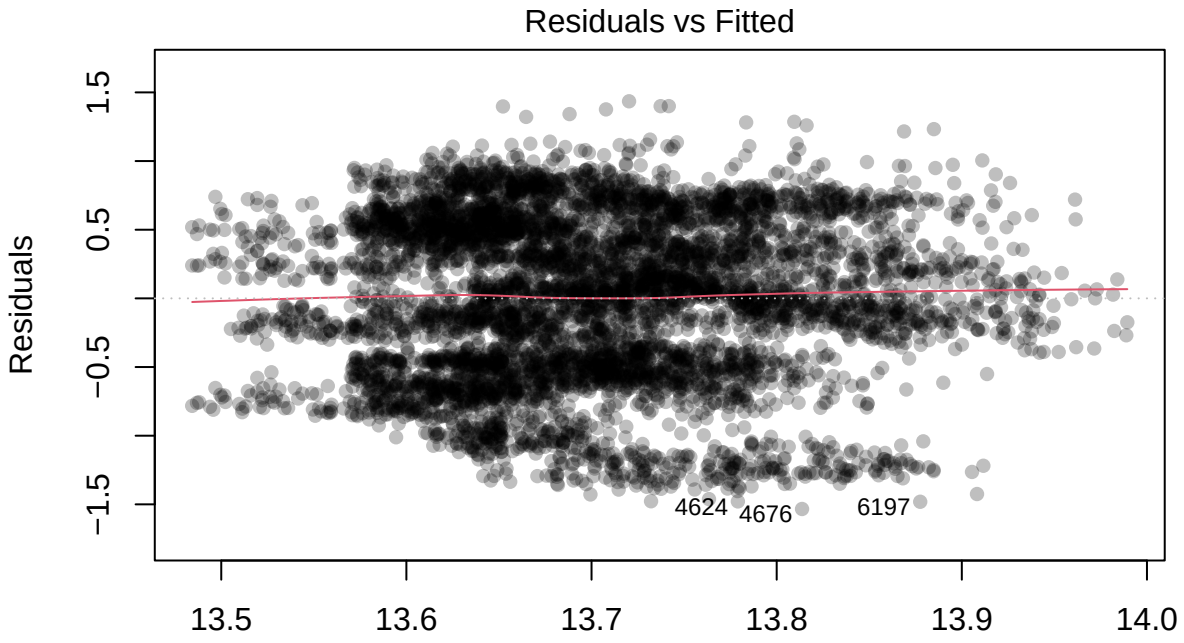

Residuals vs Fitted
Residuals
3e+06
1e+06
−1e+06
1906
2764
1334
900000
1000000
1100000
1200000
1300000
Fitted values
lm(Weekly_Sales ~ Days_since + Unemployment + CPI + Fuel_Price + Temperatur .

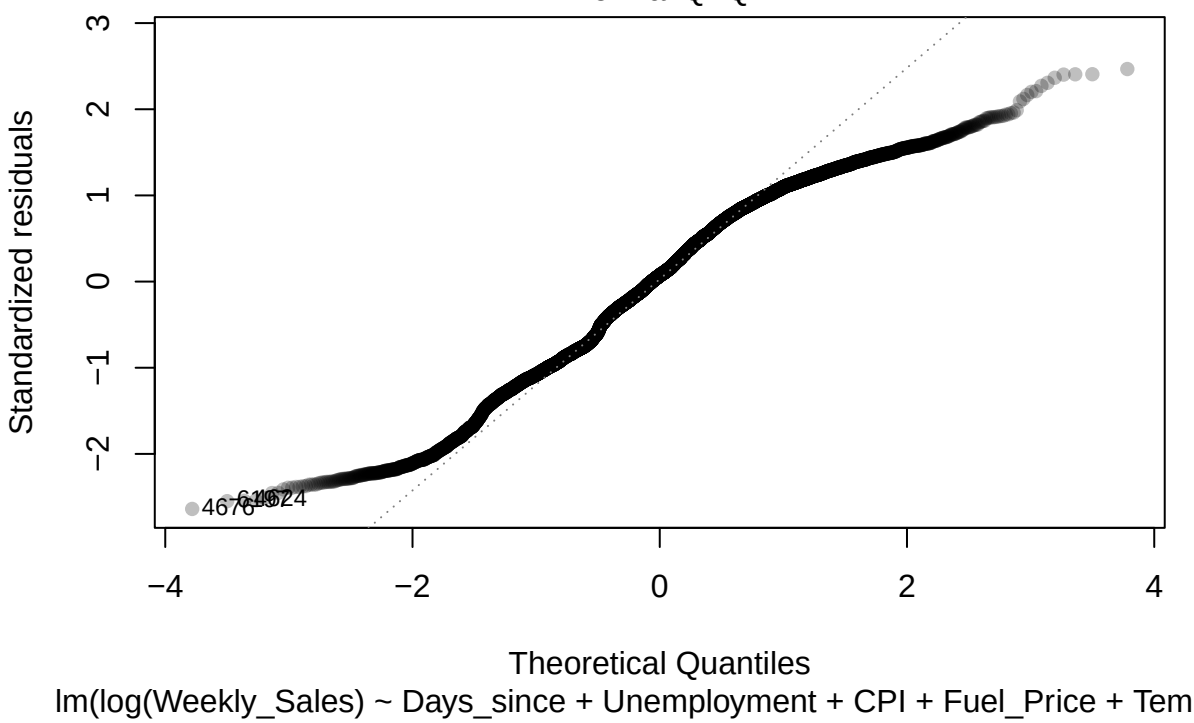

Normal Q−Q
Standardized residuals
4
2
0
−2
1906
2764
1334
−4
−2
0
2
4
Theoretical Quantiles
lm(Weekly_Sales ~ Days_since + Unemployment + CPI + Fuel_Price + Temperatur .

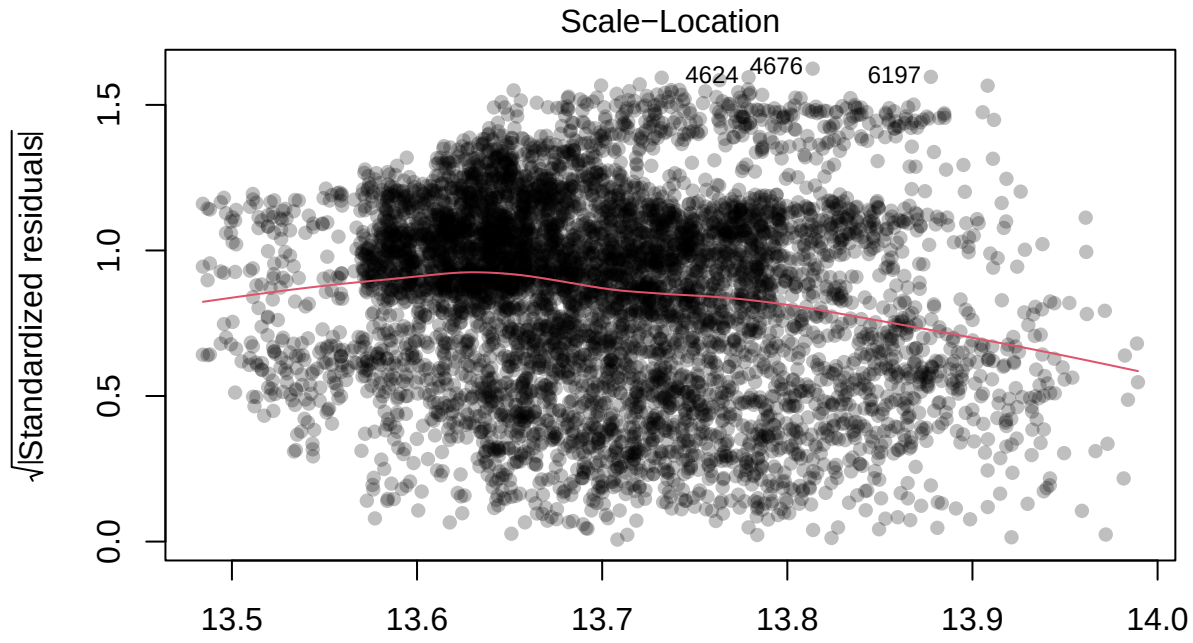


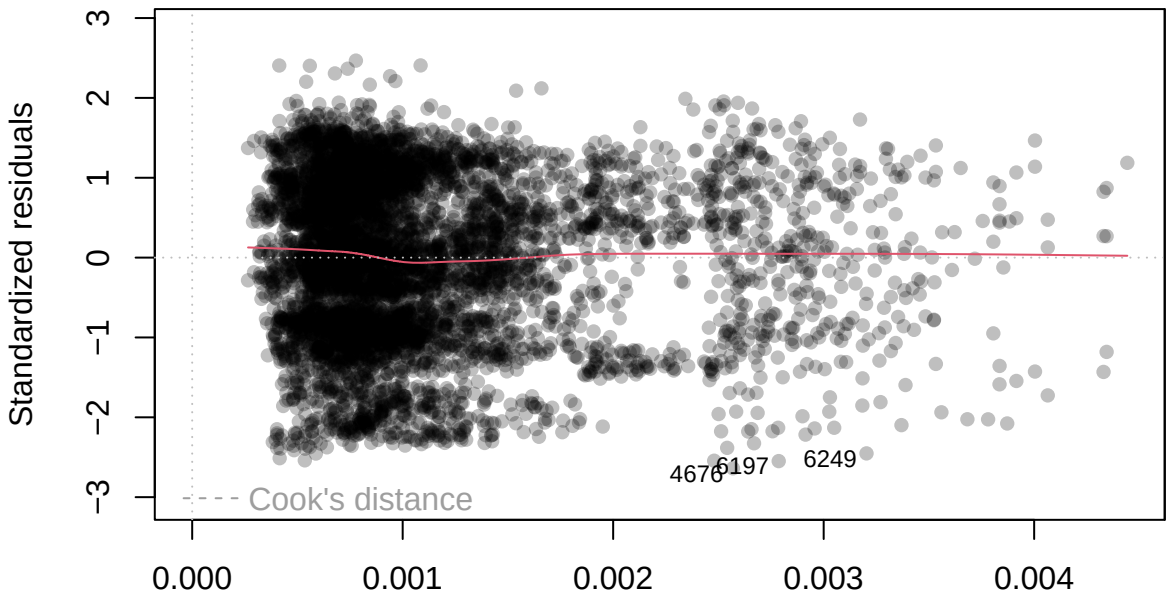



```
full_mod_w_log <- lm(log(Weekly_Sales) ~ Days_since + Unemployment + CPI + Fuel_Price + Temperature + H
plot(full_mod_w_log,
     col = rgb(red = 0, green = 0, blue = 0, alpha = 0.25),
     pch = 16
)
```

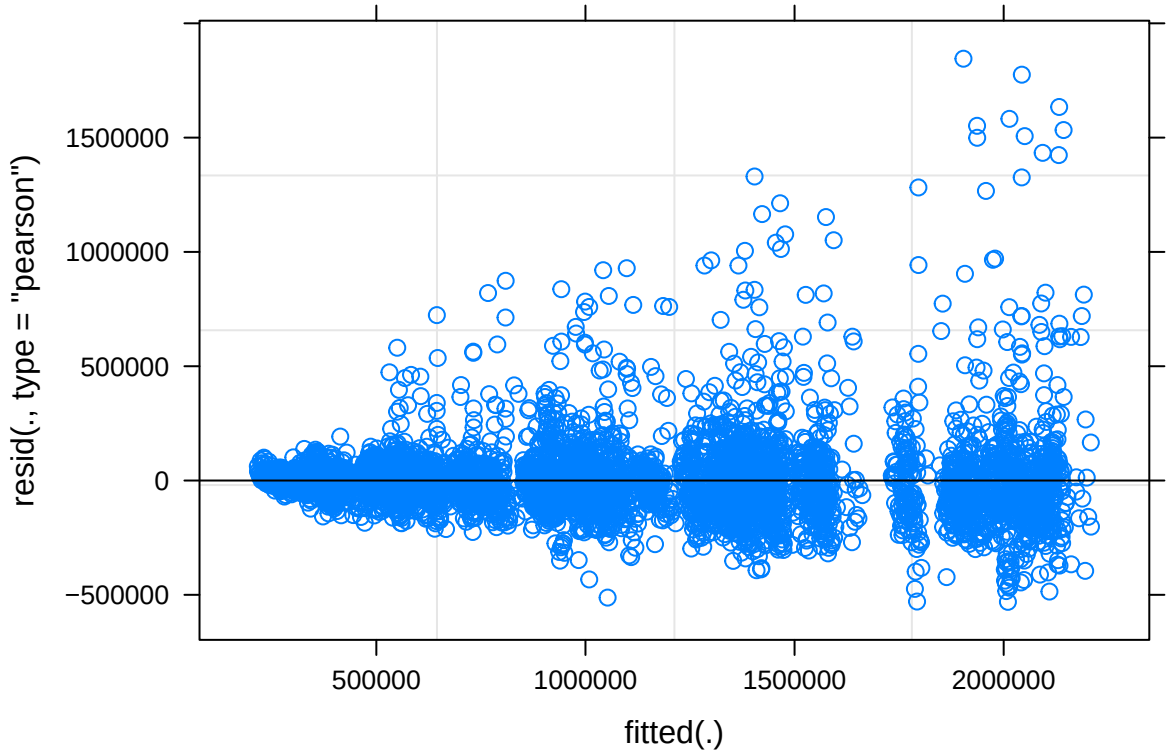

Residuals vs Fitted
Residuals
1.5
0.5
−0.5
−1.5
4624
4676
6197
13.5
13.6
13.7
13.8
13.9
14.0
Fitted values
lm(log(Weekly_Sales) ~ Days_since + Unemployment + CPI + Fuel_Price + Tempe .


Normal Q−Q
Standardized residuals
3
2
1
0
−1
−2
4676
−4
−2
0
2
4
Theoretical Quantiles
lm(log(Weekly_Sales) ~ Days_since + Unemployment + CPI + Fuel_Price + Tempe .

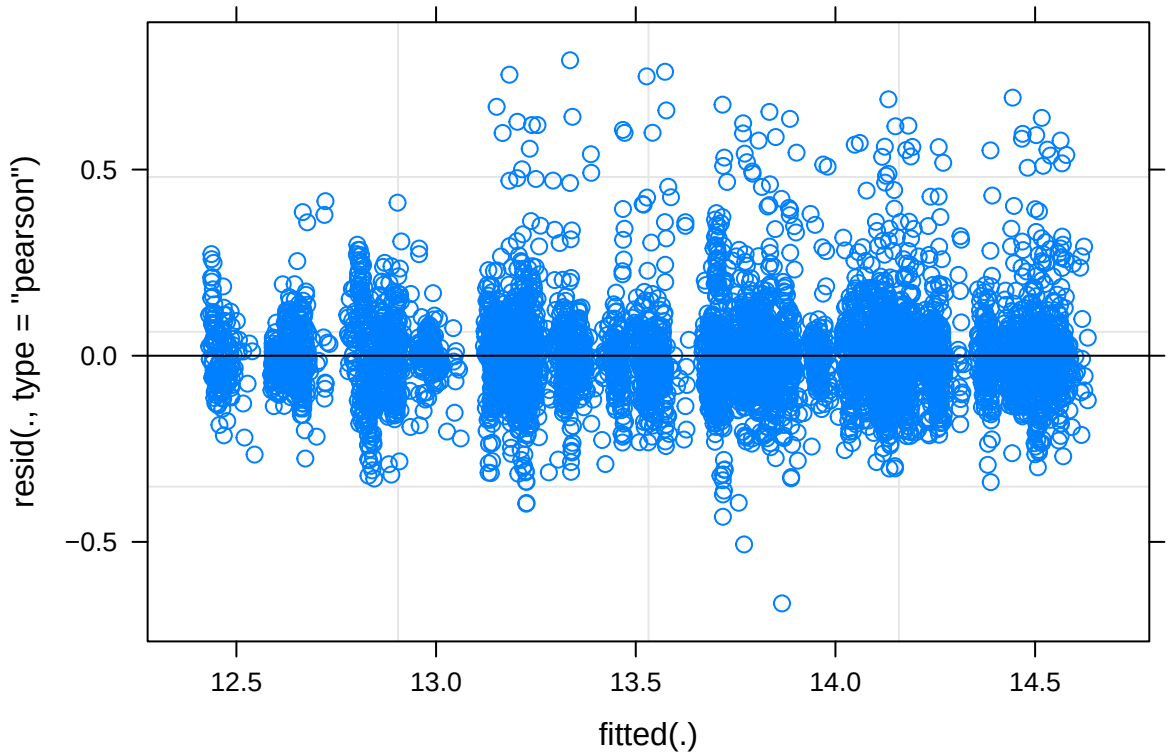




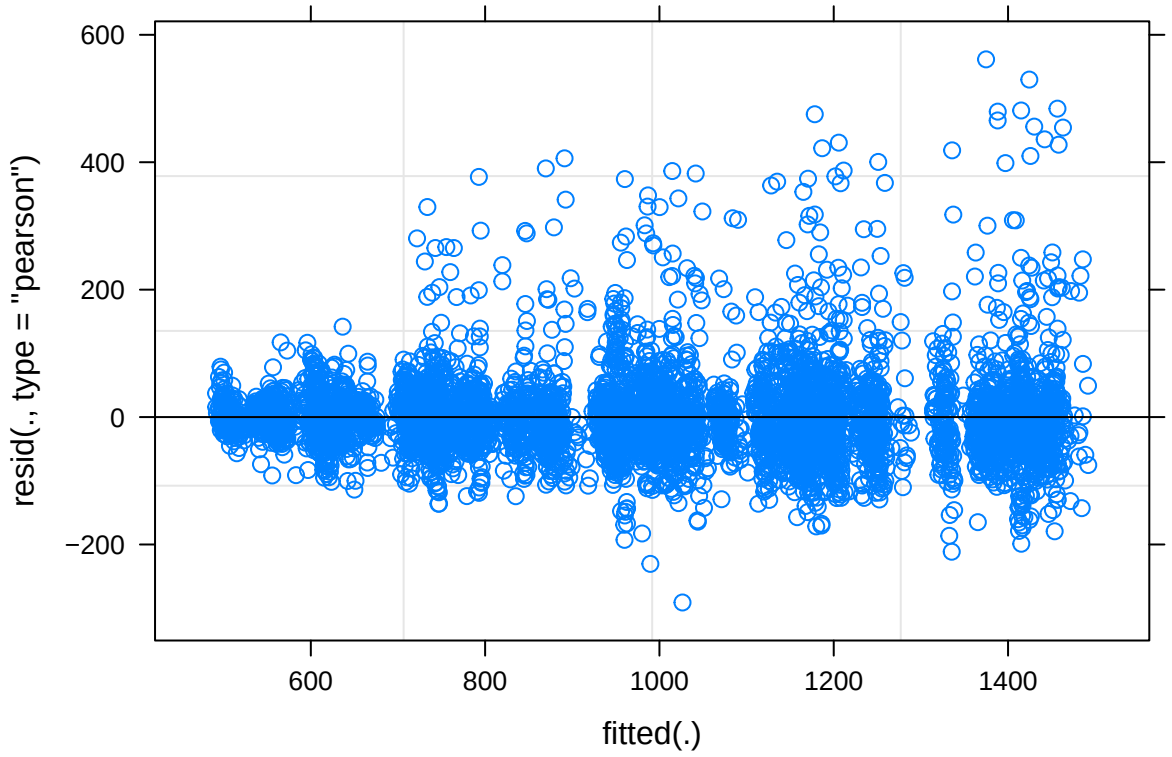

```
# Fit a mixed model, look at residual plots
mixed_effect_mod <- lmer(Weekly_Sales ~ (1|Store) + Days_since + Unemployment + CPI + Fuel_Price + Tempe
plot(mixed_effect_mod)
```




```
mixed_effect_mod_log <- lmer(log(Weekly_Sales) ~ (1|Store) + Days_since + Unemployment + CPI + Fuel_Pri
plot(mixed_effect_mod_log)
```

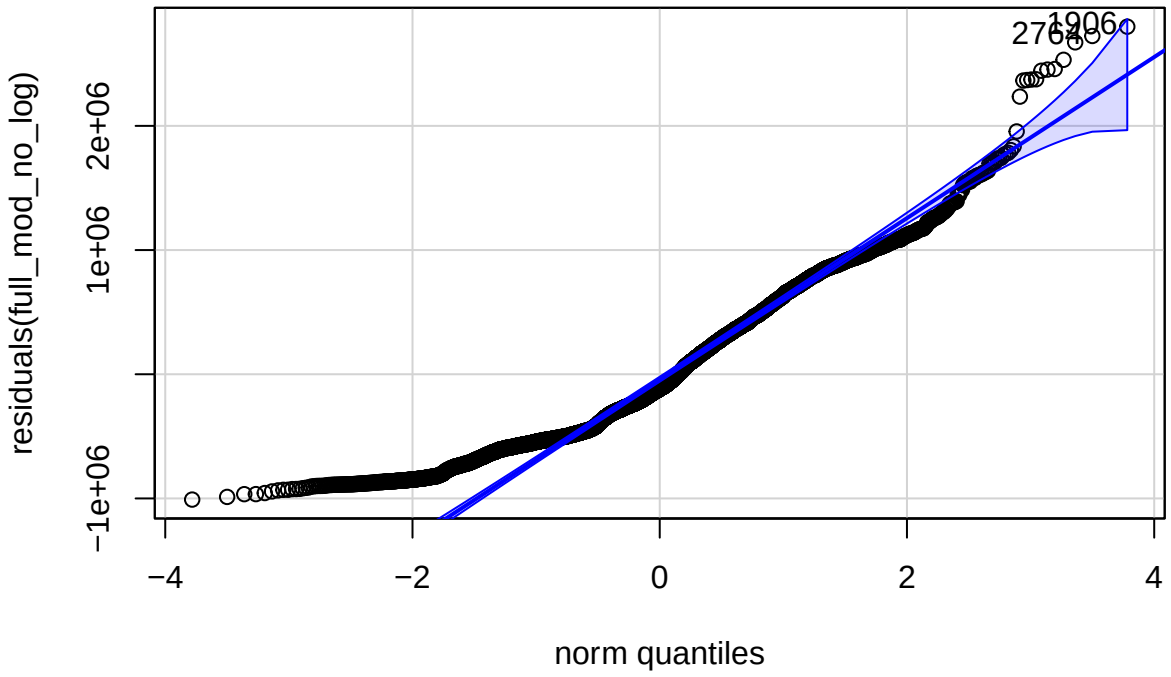


```
mixed_effect_mod_sqrt<- lmer(sqrt(Weekly_Sales) ~ (1|Store) + Days_since + Unemployment + CPI + Fuel_Pr
plot(mixed_effect_mod_sqrt)
```



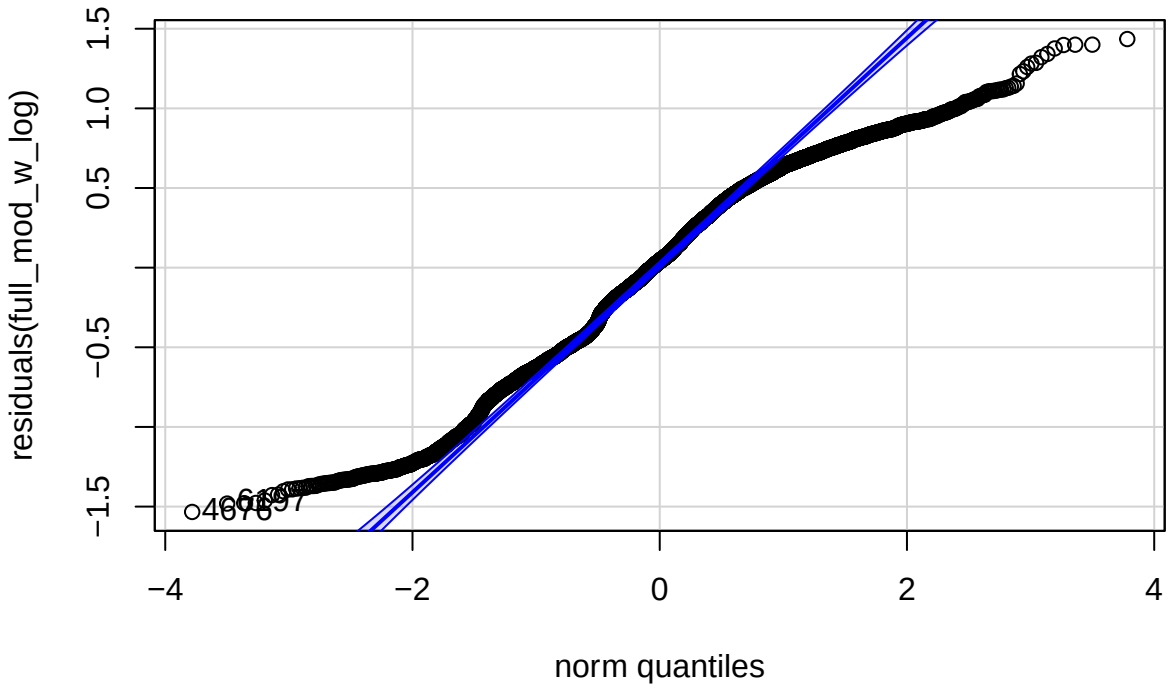
No transformation on Weekly Sales and sqrt(Weekly Sales) show a megaphone effect, nonconstant variance. However, the log transformation on Weekly Sales seems to get rid of the megaphone appearance of the residuals, indicating constant variance. Thus, the log transformation may be the most appropriate.

```
# QQ plots for different models
qqPlot(residuals(full_mod_no_log))
```



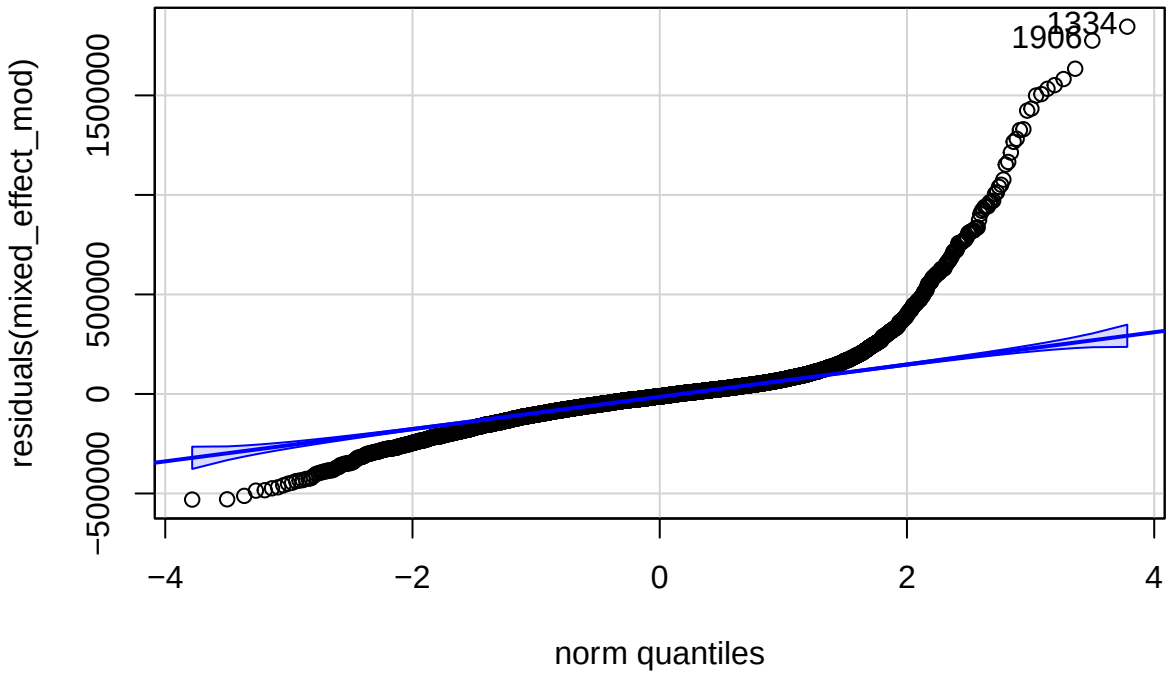

```
## [1] 1906 2764
```

```
qqPlot(residuals(full_mod_w_log))
```



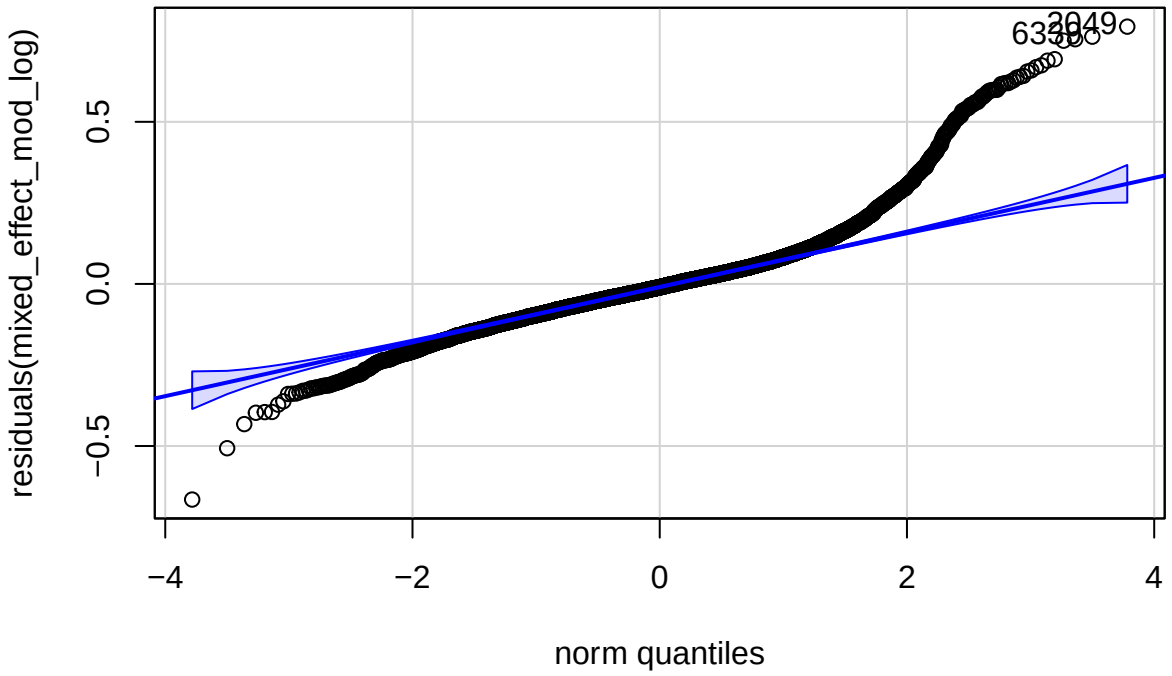

```
## [1] 4676 6197
```

```
qqPlot(residuals(mixed_effect_mod))
```




```
## [1] 1334 1906
```

```
qqPlot(residuals(mixed_effect_mod_log))
```

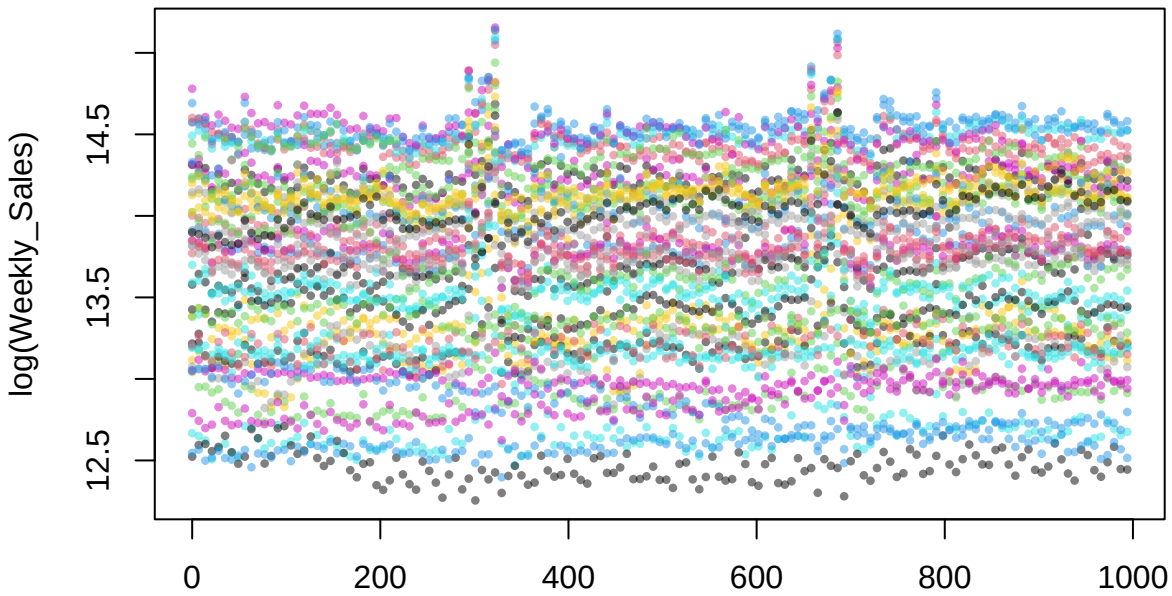


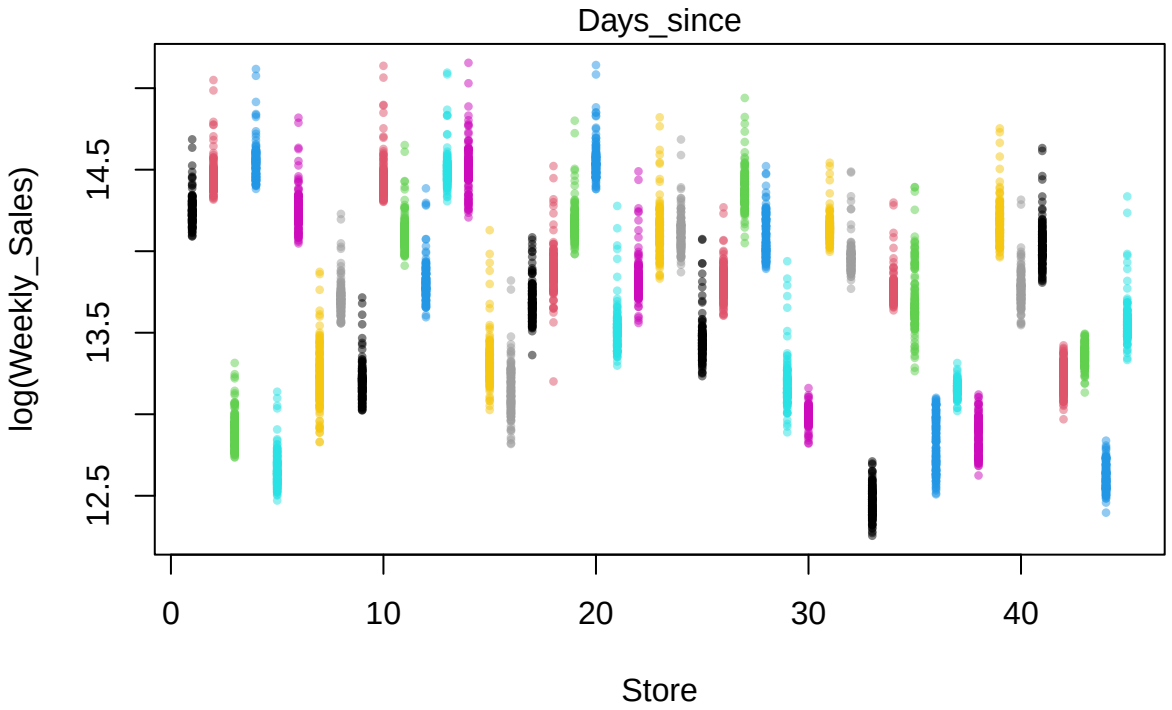

```
## [1] 2049 6339
```

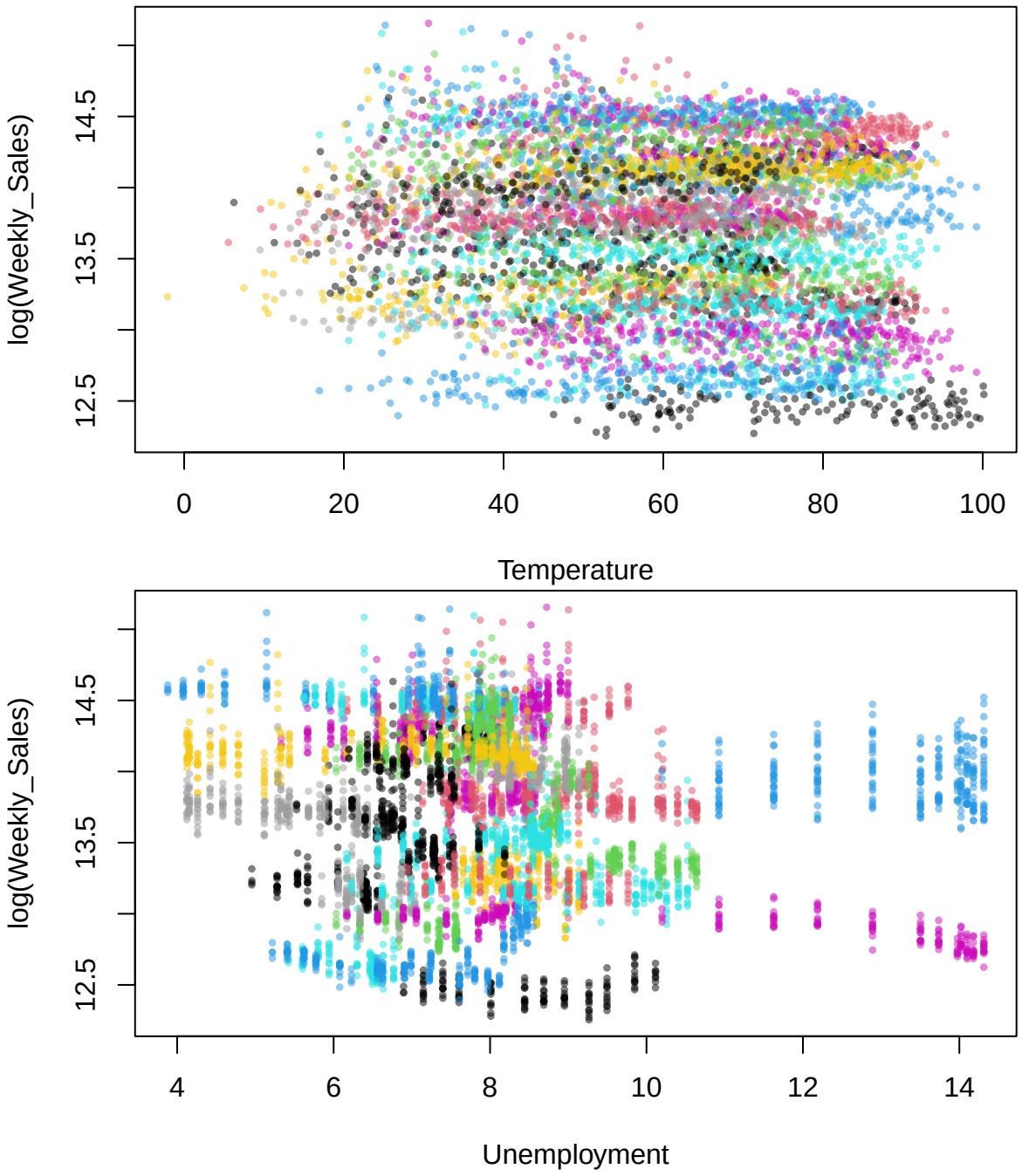
The full model without the log transform shows residuals that are most close to a normal distribution. However, all show deviations at lower normal quantiles and higher normal quantiles. When comparing the mixed effect on store on log(Weekly Sales) versus a mixed effect on store on Weekly Sales, the one with the log transformation appears to be more close to a normal distribution.

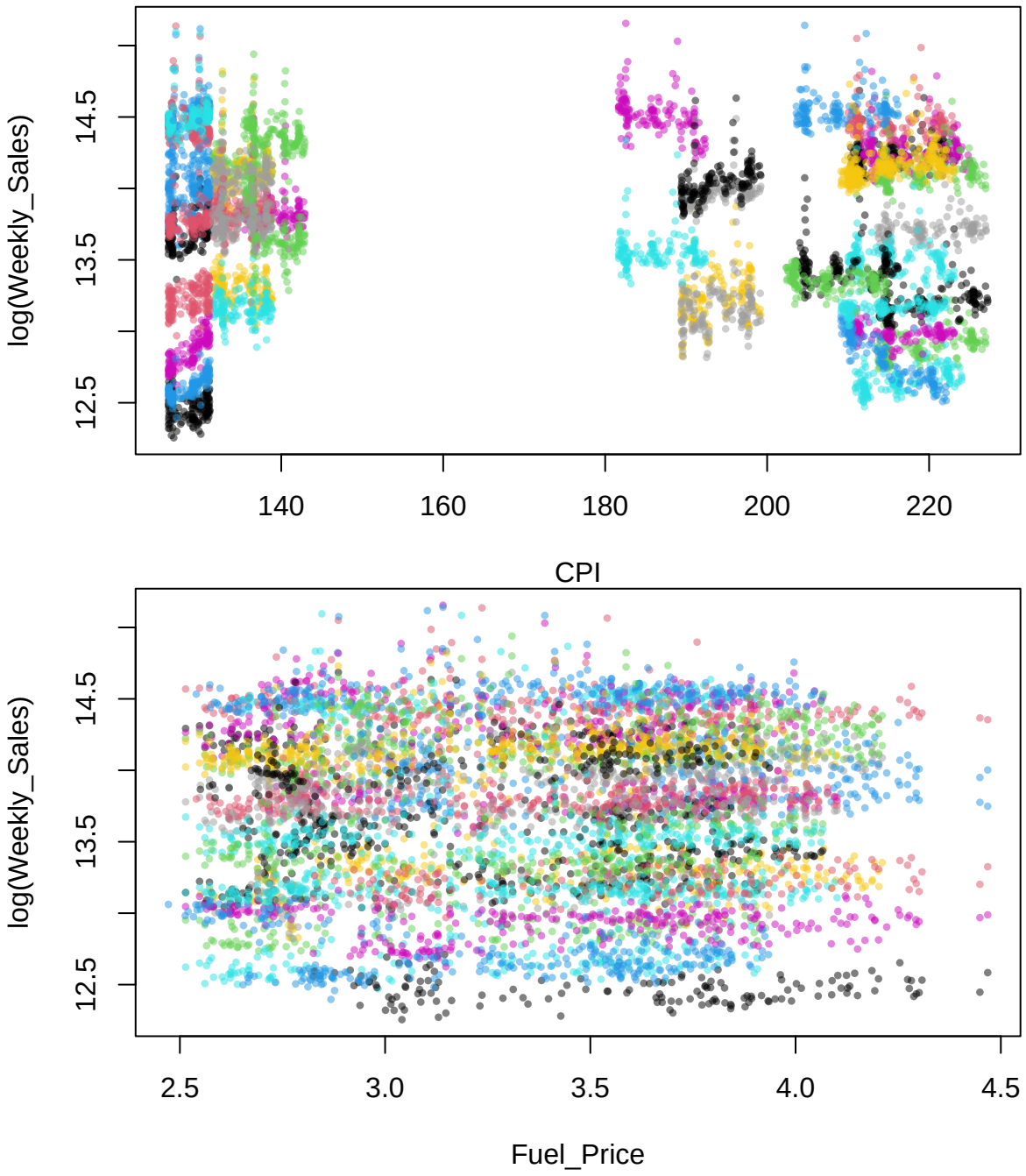
Scatterplot with variabels vs log(Weekly_Sales)

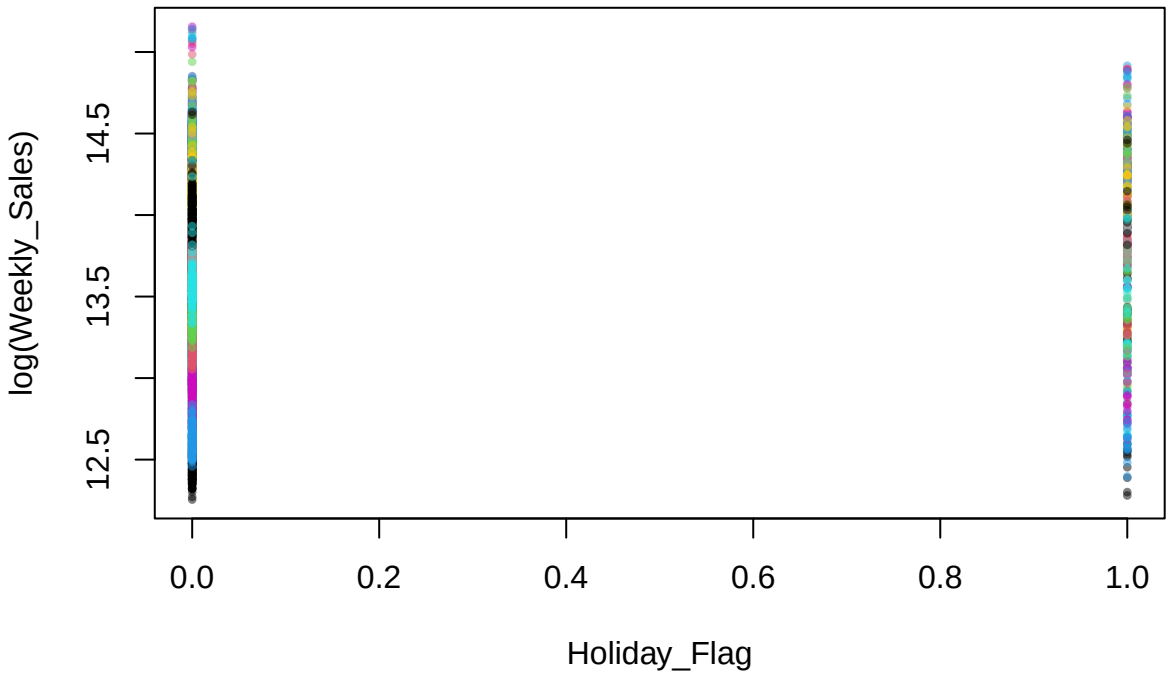
```
# Looking at effect of Days and Store on log(Weekly Sales)
plot(log(Weekly_Sales) ~ Days_since + Store + Temperature + Unemployment + CPI + Fuel_Price + Temperatu
     col = alpha(walmart$Store, 0.5),  # makes color dependent on store
     pch = 16,
     cex = 0.6)
```






log(Weekly_Sales)
12.5
13.5
14.5
0
20
40
60
80
100
Temperature
log(Weekly_Sales)
12.5
13.5
14.5
4
6
8
10
12
14
Unemployment


log(Weekly_Sales)
12.5
13.5
14.5
140
160
180
200
220
CPI
log(Weekly_Sales)
12.5
13.5
14.5
2.5
3.0
3.5
4.0
4.5
Fuel_Price




```
# Scatterplot matrix - uninterprettable
pairs(Weekly_Sales ~ Days_since + Unemployment + CPI + Fuel_Price + Temperature + Holiday_Flag, data = w
```