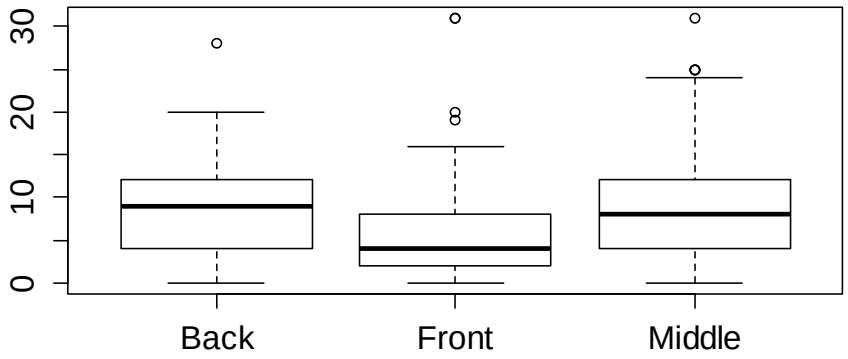
ONE-WAY ANALYSIS OF VARIANCE: Party Days per Month by Seat Location Example

There are 685 students in the dataset. Y = PartyDays (days per month the student reported that they go to parties) and there is one categorical variable, "Seat" which is a response to the question "Where do you typically sit in a classroom – in the front, middle or back?" We want to know if population mean for "Party Days" differ for students who typically sit in the 3 classroom locations. If so, we want to know which locations have means that are significantly different.

```
> attach(Student0405) #So we don't have to type the data set name each time
> tapply(PartyDays,Seat,mean) #Gives means for each group
    Back    Front   Middle
8.507463 5.295302 8.027363
> tapply(PartyDays,Seat,sd)#Gives standard deviations; they are very close
    Back    Front   Middle
5.379921 5.143224 5.386484
> tapply(PartyDays,Seat,length) #Gives sample sizes for each group
  Back  Front Middle
   134    149    402
> boxplot(PartyDays~Seat) #Get a picture of the data
```

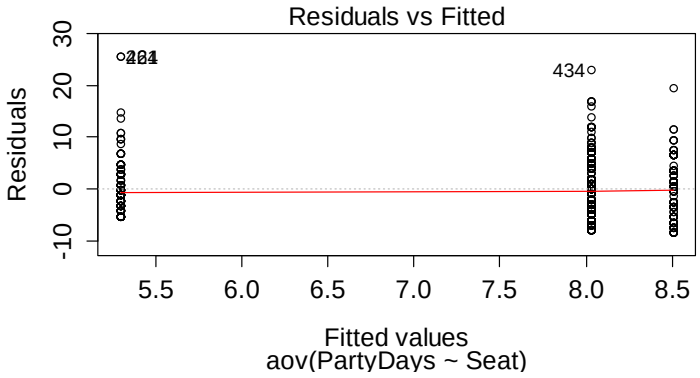


Note that the maximum answer is 31, since these are days per month. There are some outliers, but with such a large sample size that's fine. Values for the "Front" group look smaller; Back and Middle look similar.

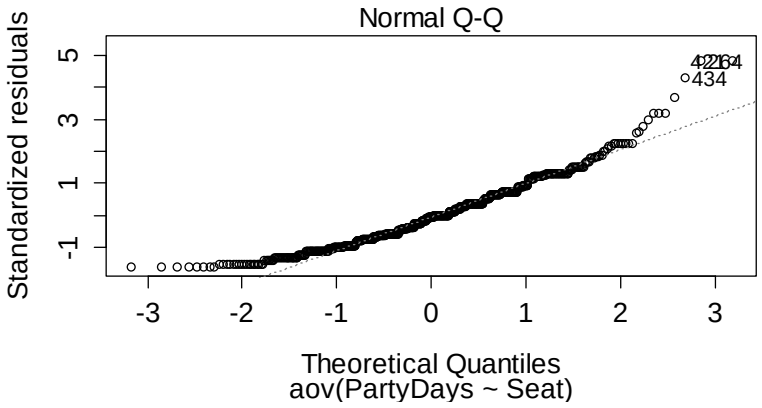
```
> Party<-aov(PartyDays~Seat) #Fit the ANOVA model
> Party   #This shows SS, DF, and the square root of MSE.
Call:
   aov(formula = PartyDays ~ Seat)
Terms:
                     Seat Residuals
Sum of Squares    971.552 19399.198
Deg. of Freedom         2       682

Residual standard error: 5.333345
Estimated effects may be unbalanced
> summary(Party) #This provides the usual ANOVA Table
             Df Sum Sq Mean Sq F value   Pr(>F)
Seat          2    972   485.8   17.08 5.79e-08 ***
Residuals   682  19399    28.4
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Note that we can reject H_0: Population mean party days are the same for students in the 3 seat locations. Let's check the conditions, then find out which means are significantly different.

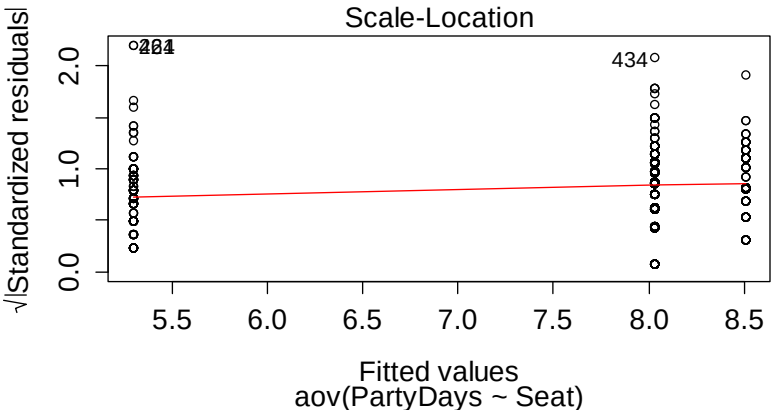
```
> plot(Party)  #Provides plots you can cycle through. Here they are.
```



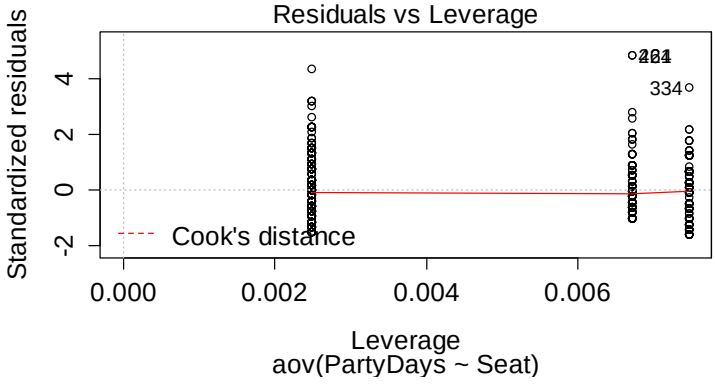
The "Fitted values" are the sample means for the three seat locations; the Residuals are for all individuals, but in this plot they are not standardized.



It looks like there is some skewness at the upper end, but with a large data set like this one that doesn't matter.

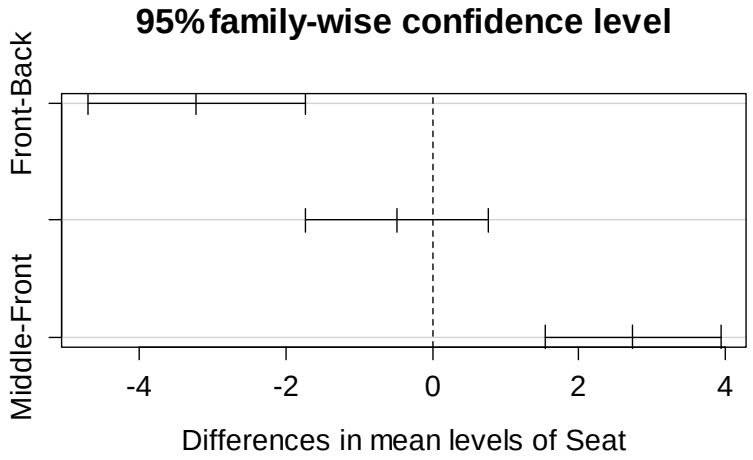


This plot has the positive square root of the standardized residuals. There are some large ones, for people who partied much more than others. These represent natural variability, so they should not be removed. With such a large sample they are not a problem.



Leverage has no useful meaning when the explanatory variable is categorical.

```
> #If you want the test statistic and p-value only
> oneway.test(PartyDays~Seat,var.equal=T)
     One-way analysis of means

data:  PartyDays and Seat
F = 17.078, num df = 2, denom df = 682, p-value = 5.793e-08

> TukeyHSD(Party,ordered=T) #Get Tukey CIs with means ordered
  Tukey multiple comparisons of means
    95% family-wise confidence level
    factor levels have been ordered

Fit: aov(formula = PartyDays ~ Seat)

$Seat
                   diff        lwr      upr     p adj
Middle-Front 2.7320612  1.5305652 3.933557 0.0000004
Back-Front   3.2121607  1.7207392 4.703582 0.0000016
Back-Middle  0.4800995 -0.7694966 1.729696 0.6389475

> #Note that Front is significantly different from Middle and Back
> plot(TukeyHSD(Party))
```



Note that the only interval that covers 0 is the one for Middle – Back (not labeled), again verifying that the populations mean party days differ for Front and Back, and for Front and Middle, but we cannot conclude that they differ for Middle and Back.

We can also do pairwise t-tests for all pairs of means. The results show (again) that the population means for Middle and Back are not significantly different, but the other pairs are:

```
> pairwise.t.test(PartyDays,Seat,p.adj="none")

     Pairwise comparisons using t tests with pooled SD

data:  PartyDays and Seat

       Back    Front
Front  5.4e-07 -
Middle 0.37    1.3e-07

P value adjustment method: none
```

The LSD method can be done “by hand” by constructing the intervals using individual R commands, but it’s better to use the Tukey method.