One-Way ANOVA Solutions

5.22 Child Poverty

a. The null hypothesis is that all three sizes of counties have the same mean child poverty rate. The alternative is that at least one of the three types of counties has a different mean child poverty rate than the others.

In symbols:
$H_0 : \alpha_1 = \alpha_2 = \alpha_3 = 0$ or $H_0 : \mu_1 = \mu_2 = \mu_3$
H_a : at least one $\alpha_k \neq 0$ or H_a : at least one μ is different than the others.

b. The dotplot suggests that there may be very different amounts of variability between the 3 types of county. In particular, it appears that small counties have very little variability with respect to child poverty rates compared to medium and large counties. Therefore, we are concerned about the equal variances condition.

5.24 Fantasy Baseball

a.

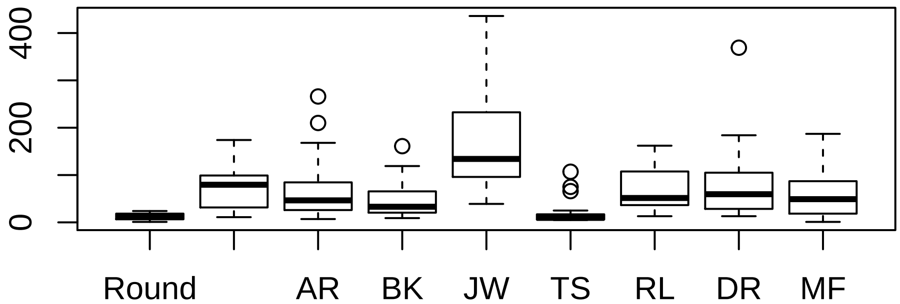
```
#Putting the data in a more useful form (I gave you this code):
library(reshape2)
Baseball2 <- melt(FantasyBaseball[,2:9])
```

```
## No id variables; using all as measure variables
```

```
Baseball2$Round <- rep(1:24) #add the Round variable
colnames(Baseball2)[1] <- c("Person"); colnames(Baseball2)[2] <- c("Time") #make the column names usefu
favstats(Time~Person, data=Baseball2)
```

```
##   Person min    Q1 median     Q3 max      mean        sd  n missing
## 1     DJ  11 32.25   79.5  99.00 174  69.62500  41.61502 24       0
## 2     AR   7 26.00   46.5  79.25 266  68.29167  66.89153 24       0
## 3     BK   9 20.75   33.0  65.25 161  47.95833  39.25112 24       0
## 4     JW  39 98.50  134.0 231.25 436 163.87500 104.16555 24       0
## 5     TS   5  6.00    9.5  17.00 107  19.33333  25.82999 24       0
## 6     RL  13 36.75   51.5 105.25 162  67.12500  44.55120 24       0
## 7     DR  13 29.75   59.5 103.50 369  80.12500  75.83467 24       0
## 8     MF   1 18.75   49.0  86.50 187  63.83333  55.99664 24       0
```

```r
boxplot(FantasyBaseball)
```



The boxplots show that most of the distributions are skewed to the right. This makes sense since some decisions are harder and take longer to make. Most participants take about the same amount of time to decide, but JW is quite a bit slower and TS is faster than the rest.

b. In order to run the ANOVA, the data must be "unstacked", which you can do use `melt` from the `reshape2` package:

```r
model5.24 <- aov(Time~Person, data=Baseball2); Anova(model5.24)
```

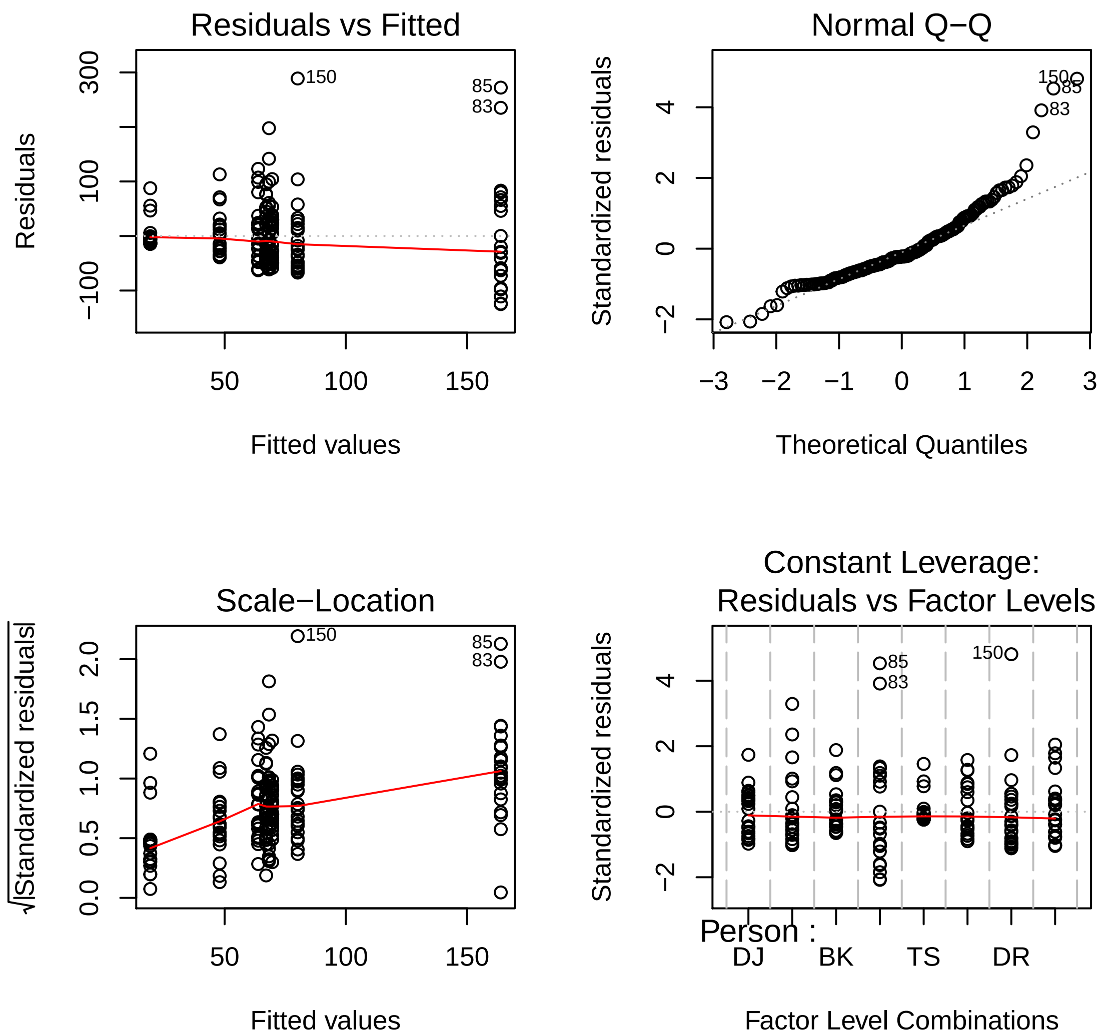
```
## Anova Table (Type II tests)
##
## Response: Time
##           Sum Sq  Df F value    Pr(>F)
## Person    287196   7  10.891 1.788e-11 ***
## Residuals 693126 184
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The very small p-value provides strong evidence that at least one of the participants has a different mean selection time than the others.

5.25 Fantasy Baseball (continued)

a.

```r
par(mfrow=c(2,2))
plot(model5.24)
```



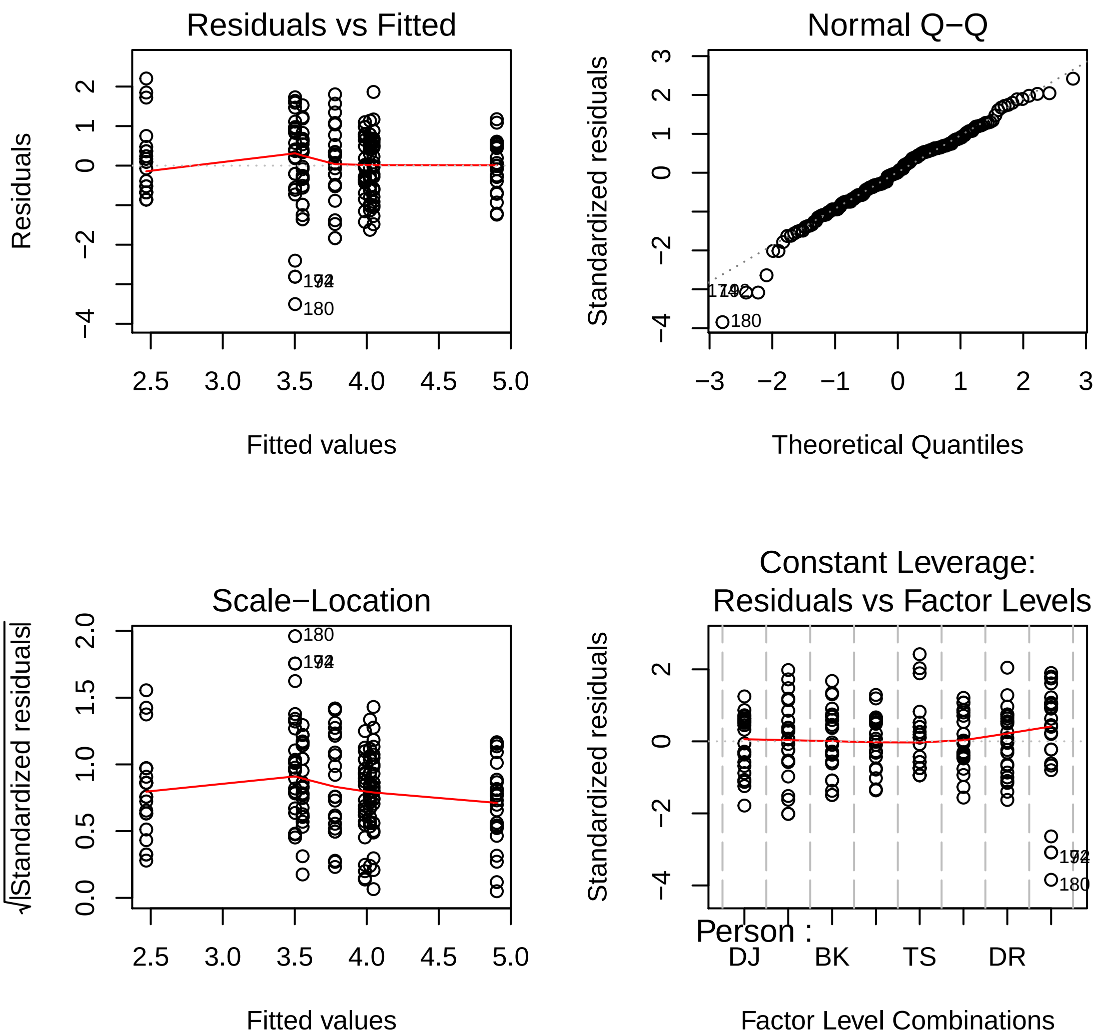
Clearly, the residuals do not come from a population with a normal distribution, since the plot curves a lot in the upper quantiles. Additionally, the assumption of constant variance is questionable, since the resids vs. fitted plot shows a megaphone pattern, and s(max)/s(min) = 104.17/25.83 = 4.03 (which is greater than 2). The conditons of the ANOVA model are not met.

b.

```
Baseball2$lnTime <- log(Baseball2$Time)
model5.25 <- aov(lnTime~Person, data=Baseball2); Anova(model5.25)
```

```
## Anova Table (Type II tests)
##
## Response: lnTime
##           Sum Sq  Df F value    Pr(>F)
## Person     78.75   7  12.989 1.538e-13 ***
## Residuals 159.37 184
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
par(mfrow=c(2,2))
plot(model5.25)
```



Both normality and constant variance are much improved when we use ln(Time) as the response variable; the conditions of the ANOVA model are now met to my satisfaction.

The very small p-value provides strong evidence that at least one of the participants has a different mean log selection time than the others.

5.27 Fantasy Baseball (continued)

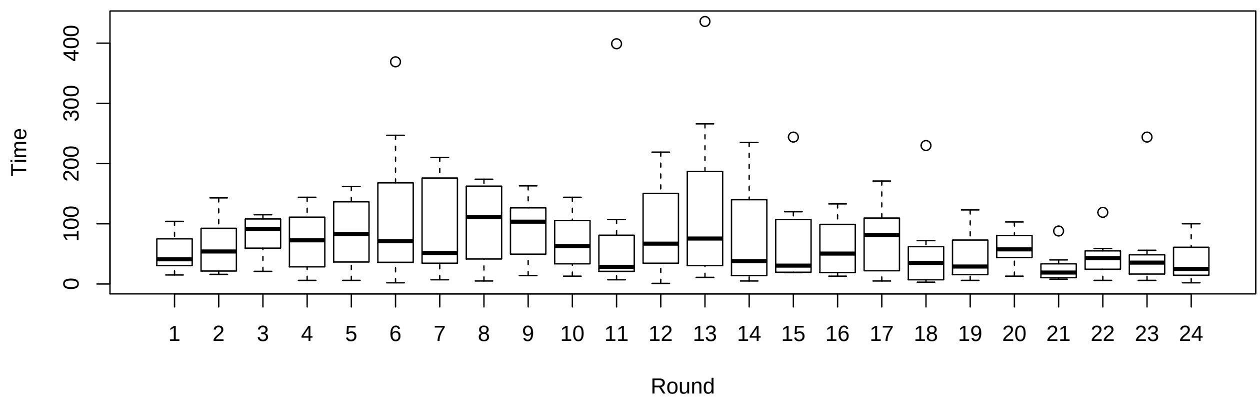
Let's investigate the Time values by Round using EDA first. The boxplots show that most rounds are right-skewed, and several have outliers. It is not obvious that any particular Round is significantly longer or shorter than any other, although the middle rounds do seem to be longer (and have more outliers) than the very early rounds (1 and 2) or later rounds (18+).

```
favstats(Time~Round, data=Baseball2)
```

```
##   Round min    Q1 median     Q3 max   mean       sd n missing
## 1     1  15 32.75   41.0  62.00 104 51.500 33.14470 8       0
## 2     2  16 23.75   54.0  88.25 143 61.875 45.31300 8       0
## 3     3  21 62.75   91.5 105.50 115 81.750 32.76213 8       0
## 4     4   6 34.75   72.5 105.00 144 71.750 49.35513 8       0
## 5     5   6 42.25   83.0 132.75 162 85.000 58.99637 8       0
```

```
## 6       6   2 45.50    71.0 128.50 369 115.125 126.84235 8       0
## 7       7   7 35.75    51.5 172.00 210  92.625  80.44153 8       0
## 8       8   5 44.25   111.0 161.75 174 101.125  66.45393 8       0
## 9       9  14 61.75   103.5 122.25 163  92.000  51.73007 8       0
## 10     10  13 39.75    63.0 102.25 144  70.125  45.39175 8       0
## 11     11   7 23.00    28.5  68.00 399  83.375 131.28806 8       0
## 12     12   1 48.25    67.0 120.75 219  90.500  82.54869 8       0
## 13     13  11 32.25    75.5 147.50 436 129.125 147.61769 8       0
## 14     14   5 17.50    38.0 116.50 235  78.000  87.85052 8       0
## 15     15  19 19.75    30.5 100.50 244  72.125  79.44349 8       0
## 16     16  13 22.00    50.5  94.50 133  60.375  45.63187 8       0
## 17     17   5 28.50    81.5 102.25 171  75.250  57.72039 8       0
## 18     18   3  8.00    35.0  57.00 230  55.125  74.80534 8       0
## 19     19   6 16.75    29.0  57.00 123  45.500  43.91876 8       0
## 20     20  13 47.00    57.5  77.75 103  60.000  28.17294 8       0
## 21     21   8 10.75    19.0  30.25  88  27.750  26.51549 8       0
## 22     22   6 31.75    43.0  53.00 119  46.250  34.85378 8       0
## 23     23   6 18.25    35.5  44.75 244  56.375  77.52039 8       0
## 24     24   2 16.25    25.0  46.00 100  37.875  36.78679 8       0
```

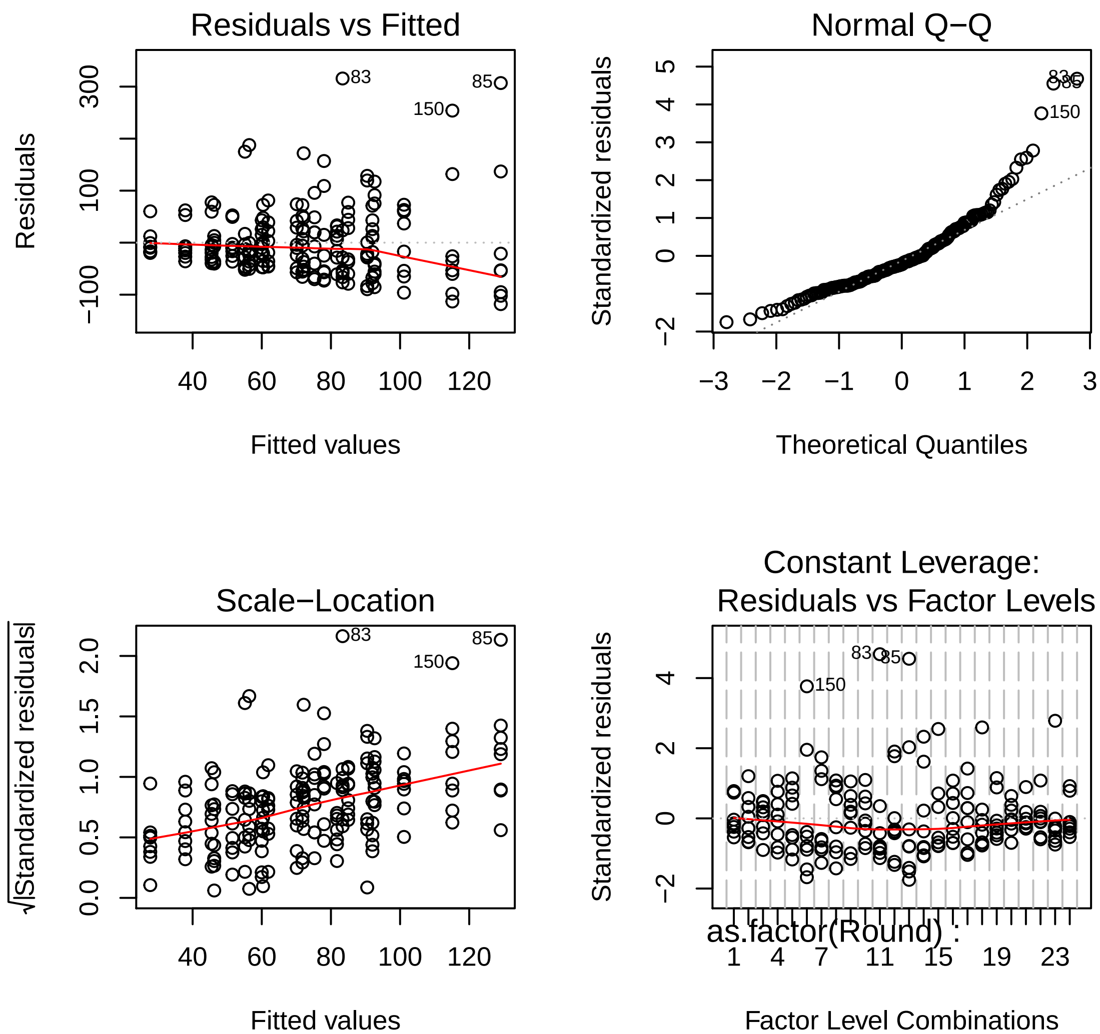
```
boxplot(Time~Round, data=Baseball2)
```




```
model5.27 <- aov(Time~as.factor(Round), data=Baseball2); Anova(model5.27) #You must use as.factor() her
```

```
## Anova Table (Type II tests)
##
## Response: Time
##                  Sum Sq  Df F value Pr(>F)
## as.factor(Round) 107158  23  0.8964 0.6033
## Residuals        873164 168
```

```
par(mfrow=c(2,2))
plot(model5.27)
```



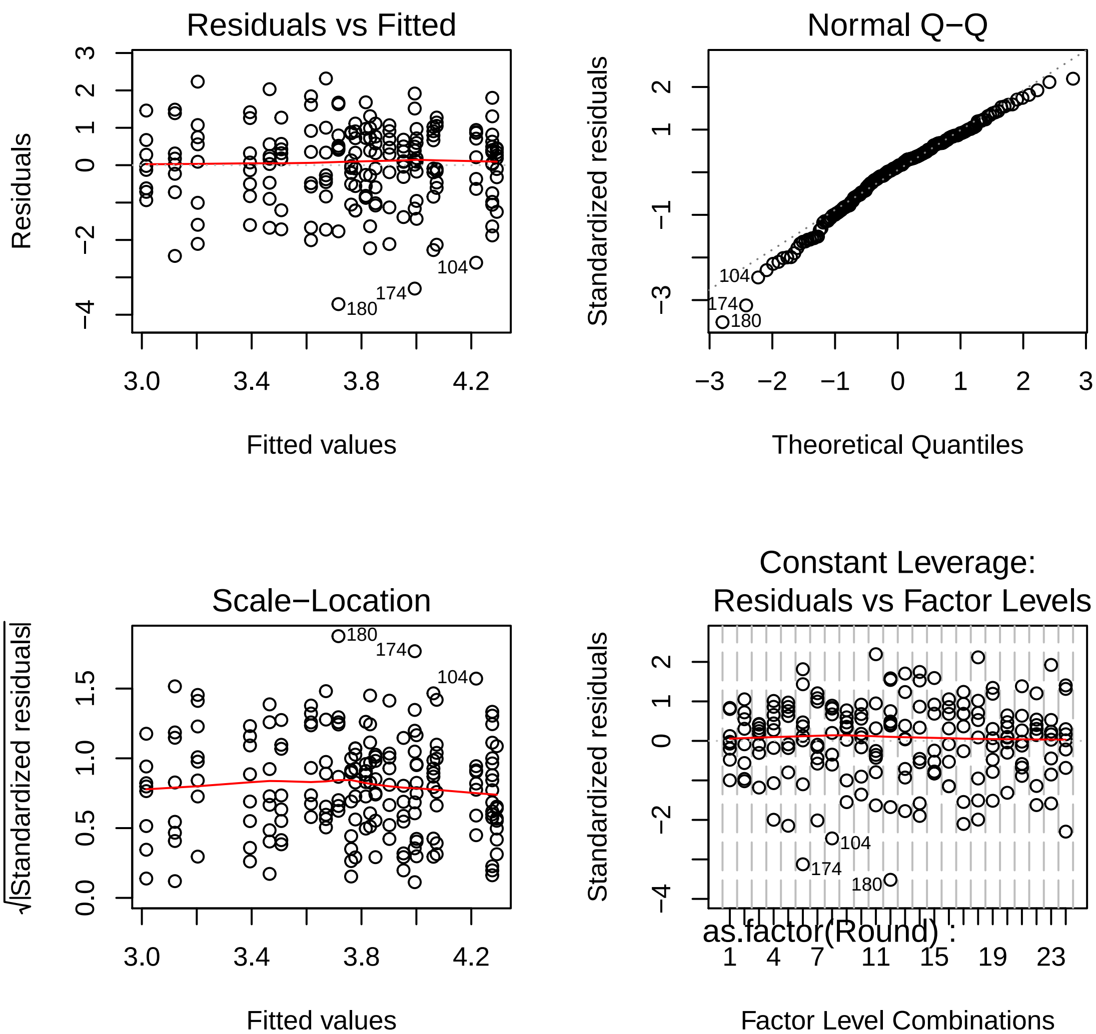
The obvious curvature in the normal probability plot shows that the residuals do not meet the normality condition. Also, there is a fan shape in the residual plot, indicating a problem with the equal variance condition. This is further supported by taking $s(\max)/s(\min) = 147.61/26.515 = 5.567$, which is much bigger than 2.

In an attempt to deal with these problems, we try using the natural log of the selection times, as we did in 5.25.

```
model5.27b <- aov(lnTime~as.factor(Round), data=Baseball2); Anova(model5.27b)
```

```
## Anova Table (Type II tests)
##
## Response: lnTime
##                   Sum Sq  Df F value Pr(>F)
## as.factor(Round)  23.823  23   0.812 0.7132
## Residuals        214.299 168
```

```
par(mfrow=c(2,2))
plot(model5.27b)
```



Now the graphs of the residuals *look* like all conditions are reasonably met. However, we should note that s(max)/s(min) = 1.839/0.571 = 3.219, which is bigger than 2. So we may still have an issue with the equal variances condition. However, since a ratio of 3.2 is much better than 5.6, we'll proceed with caution. . .

With such a large p-value, we do not have enough evidence to conclude that log of selection time differs by Round, on average.

(You could also take a square-root transformation instead of the log, which actually results in a slightly better ratio of standard deviations.)