# STA 610 HW 1

William Tirone

2023-09-15

## Question 1 (Game of Thrones)

### EDA

Initially, just considering the quantiles below with respect to gender, it seems like there is a pretty obvious difference in the screentimes between `Male`, `Female`, and `Unspecified`.

```r
tapply(gotscreen$seccount, gotscreen$Gender, summary)
```

```
$Female
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
    845    2272    2803    2999    3437    6720 

$Male
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   3856    5678    6410    7096    7869   18289 

$Unspecified
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
    182     711    1215    1390    1756    4188 
```

Now finding the total screen time for all actors of a given gender, `Male` screentime far exceeds `Female` and `Unspecified`.

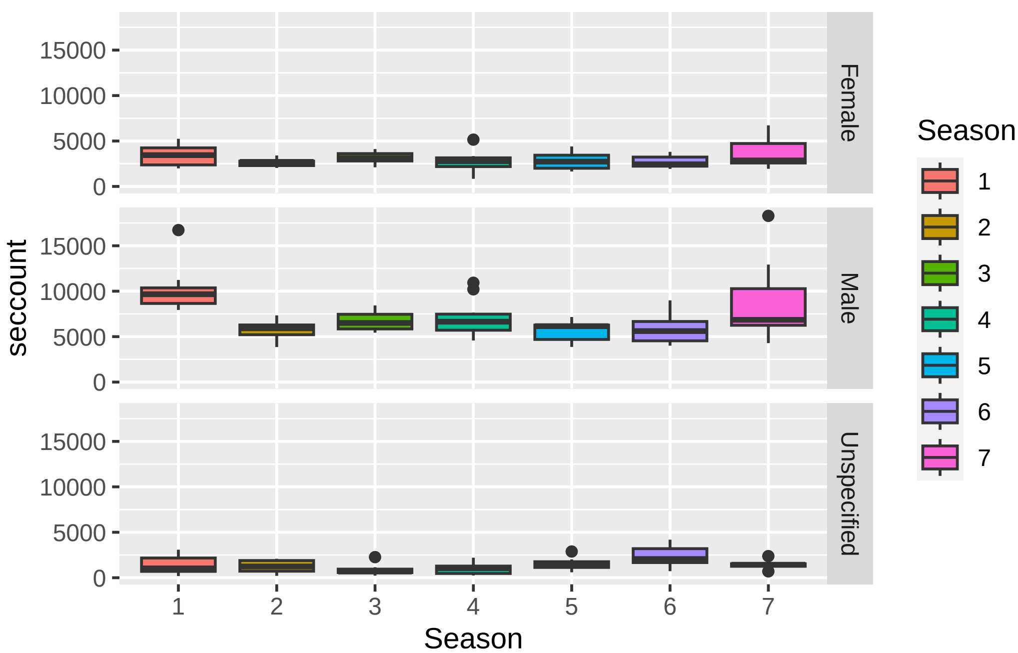
```r
gotscreen |>
  group_by(Gender) |>
  summarise(total = sum(seccount))
```

```
# A tibble: 3 x 2
  Gender      total
  <chr>       <dbl>
1 Female     200909
2 Male       475450
3 Unspecified  93128
```

And a quick plot shows that all of the maximum seconds on screen for each season are male. In fact, it looks like in seasons 1, 2, and 3, the maximum number of seconds on screen for women doesn't even exceed the minimum number of seconds on screen for men. It's also interesting that `Gender == Unspecified` is the minimum for every category. This seems like a significant omission since it represents a large chunk of the data.

The male screen time also looks a bit more variable than Female, with big changes between Season 1 & 2 then again from Season 4 to 7.

```
ggplot(gotscreen, aes(x=Season, y=seccount, group=Season)) +
    geom_boxplot(aes(fill=Season)) +
    facet_grid(Gender ~ .)
```



## Modeling and Model Fit

For data point $i$ and gender $j$ and season $k$ we can write:

$$y_{ijk} = \mu + \alpha I(j = 1) + \beta I(k = 1) + ... + \beta I(k = 6) + \gamma_{jk} + \epsilon_{ijk}$$

I'm not positive how to write the interaction effect cleanly with indicators so I have left it as a general term.

```
anova_interaction = aov(seccount ~ Gender * Season, data=gotscreen)
```

Assumptions:

1. Independence of data:

   - It seems very unlikely that the data would be independent. If a character is popular in one season, they would be much more likely to have more minutes in the next season.

2. Normally-distributed residuals

   - Condition not met, as we will see below.

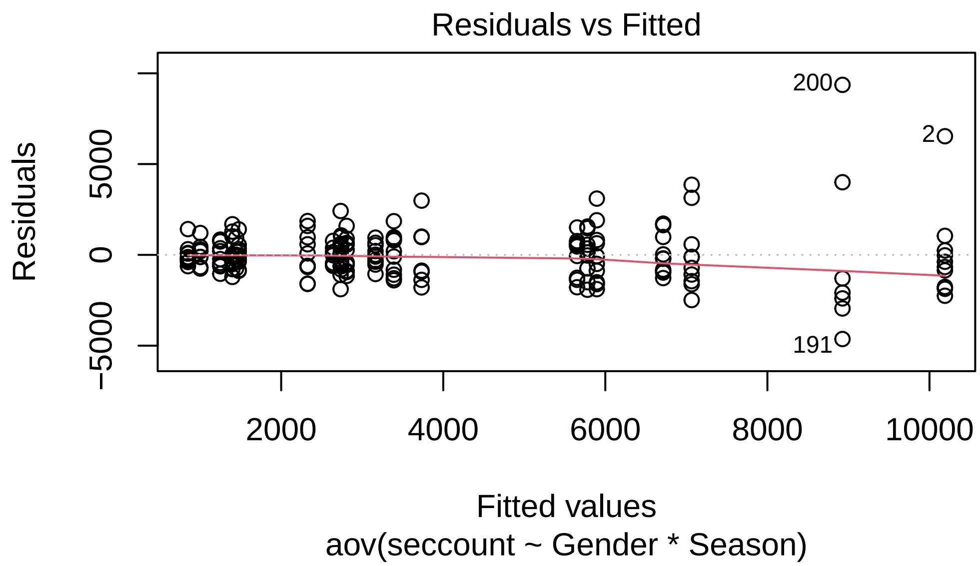
3. Homoscedasticity

   - Also not met, as discussed below.

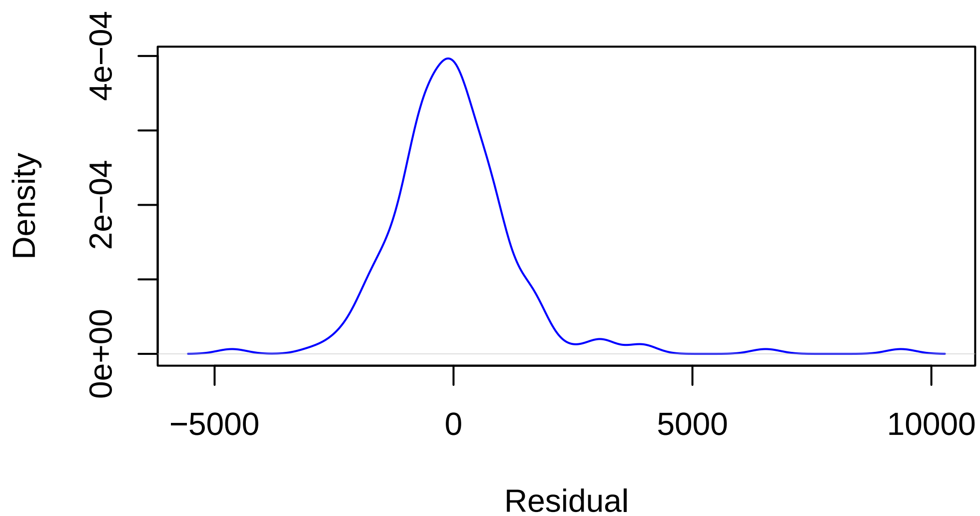
Considering the plots below, while the residuals have a (somewhat) normal-looking distribution with a heavy right tail, the Q-Q plot shows much higher values on the right tail that offers strong evidence that the errors are not actually normal.

Additionally, the Q-Q and Residuals vs Fitted plots show that observations 2, 200, and 191 are significant outliers which affect the normality and homogeneity of variance.

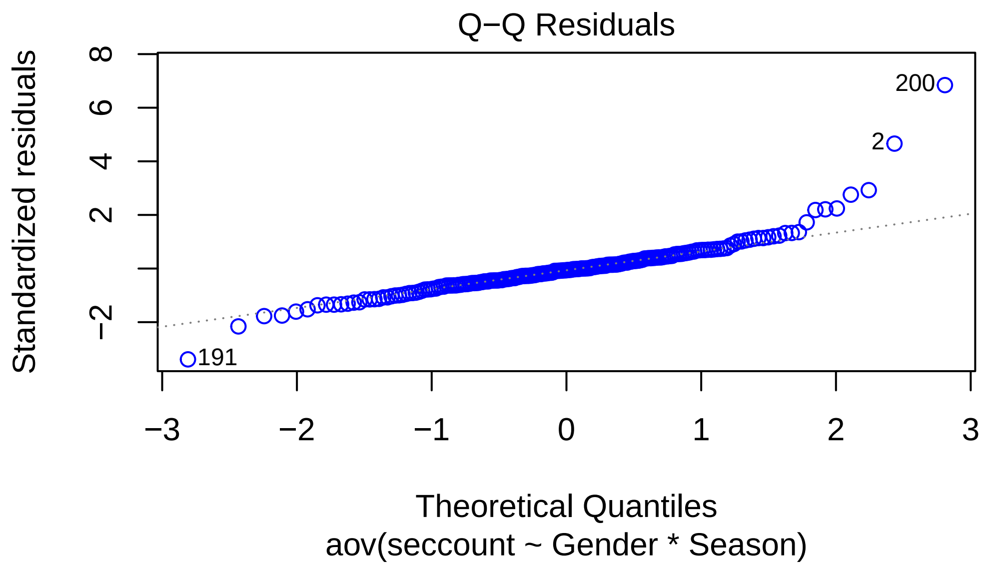
```
plot(anova_interaction, 1)
```




```
plot(density(residuals(anova_interaction)), xlab="Residual", main="", col=c("blue"))
```




```
plot(anova_interaction, which=2, col=c("blue"))
```




I also found (in a tutorial on two-way ANOVA: http://www.sthda.com/english/wiki/two-way-anova-test-in-r) the Levene test which checks for homoscedasticity. With a very small p-value in the table below, this confirms that the equal variance condition is probably not met.

I also found the Shapiro-Wilk test which checks the normality of residuals. Again, with a very small p-value, this confirms our suspicions from visually inspecting the diagnostic plots that the residuals are not normally distributed.

```
# check variances
leveneTest(seccount ~ Gender * Season, data = gotscreen)
```

```
Levene's Test for Homogeneity of Variance (center = median)
       Df F value    Pr(>F)
group  20  2.4581 0.0009218 ***
      180
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# check residuals
aov_residuals = residuals(anova_interaction)
shapiro.test(x = aov_residuals )
```

```
    Shapiro-Wilk normality test
```

```
data:  aov_residuals
W = 0.8558, p-value = 7.618e-13
```

### Parameter Estimates

Below is a table including the estimates of each parameter along with their respective confidence intervals.

```r
cbind(data.frame(anova_interaction$coefficients), confint(anova_interaction))
```

```
                            anova_interaction.coefficients      2.5 %     97.5 %
(Intercept)                                      3390.8000  2468.8259  4312.7741
GenderMale                                       6799.9000  5496.0317  8103.7683
GenderUnspecified                               -1992.6000 -3296.4683  -688.7317
Season2                                          -750.0000 -2053.8683   553.8683
Season3                                          -226.8000 -1530.6683  1077.0683
Season4                                          -656.9000 -1960.7683   646.9683
Season5                                          -583.0000 -1886.8683   720.8683
Season6                                          -650.1000 -1953.9683   653.7683
Season7                                           341.9143 -1094.8786  1778.7072
GenderMale:Season2                              -3661.1000 -5505.0483 -1817.1517
GenderUnspecified:Season2                         600.5000 -1243.4483  2444.4483
GenderMale:Season3                              -3249.7000 -5093.6483 -1405.7517
GenderUnspecified:Season3                        -321.5000 -2165.4483  1522.4483
GenderMale:Season4                              -2468.3000 -4312.2483  -624.3517
GenderUnspecified:Season4                         261.1000 -1582.8483  2105.0483
GenderMale:Season5                              -3956.0000 -5799.9483 -2112.0517
GenderUnspecified:Season5                         661.2000 -1182.7483  2505.1483
GenderMale:Season6                              -3645.7000 -5489.6483 -1801.7517
GenderUnspecified:Season6                        1578.3000  -265.6483  3422.2483
GenderMale:Season7                              -1606.3286 -3638.2606   425.6034
GenderUnspecified:Season7                        -296.1143 -2328.0463  1735.8177
```

### Conclusion

Last, we can look at the F-tests associated with the fit. These correspond to the hypothesis test:

$$H_0 = \text{all mean screen time across gender and season are the same} H_A = \text{at least two of these are not equal}$$

They all reach statistical significance with very small p-values, so we conclude that the means across gender and season are not equal. While these are convincing statistics, we have to keep in mind that the assumptions of the model were broken, so these may not be accurate results.

However, because the results from simply looking at the plots of the data agree with the test results, we could probably safely reject the null hypothesis that the means across gender and season are the same. Note, though, that this doesn't tell us which combinations are different, just that they are. To find those, we would have to conduct more testing.

```r
summary(anova_interaction)
```

```
               Df    Sum Sq   Mean Sq F value   Pr(>F)    
Gender          2 1.160e+09 579999046 265.672  < 2e-16 ***
Season          6 8.007e+07  13345037   6.113 7.54e-06 ***
Gender:Season  12 1.153e+08   9609263   4.402 3.72e-06 ***
Residuals     180 3.930e+08   2183140                     
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```