Notes on Problem Set 2.2: Understanding ANOVA

ADS2

Semester 2 2023/24

Comparing t-tests to ANOVAs

This is a question designed to make you think, so I am not going to post an answer here. But here are some directions that may help:

- What is a t-test (in essence)? What is an ANOVA (in essence)?
- If you had to do a one-way ANOVA with just two samples, what would that look like?
- Maybe you want to try it in R? Take a problem you have previously done a t-test for and run an ANOVA. What do you see? (Note that this does not constitute formal proof, but it can inform your thinking!)

Testing assumptions for ANOVAs

Here is an example of a simulated dataset with unequal variances. The other ones are quite analogous.

```
group1 <- rnorm(30, 3, 1)
group2 <- rnorm(30, 4, 0.1)
group3 <- rnorm(30, 5, 0.01)
response <- c(group1, group2, group3)
group <- c(rep(1, 30), rep(2, 30), rep(3, 30))
data <- data.frame(group, response)
head(data)
```

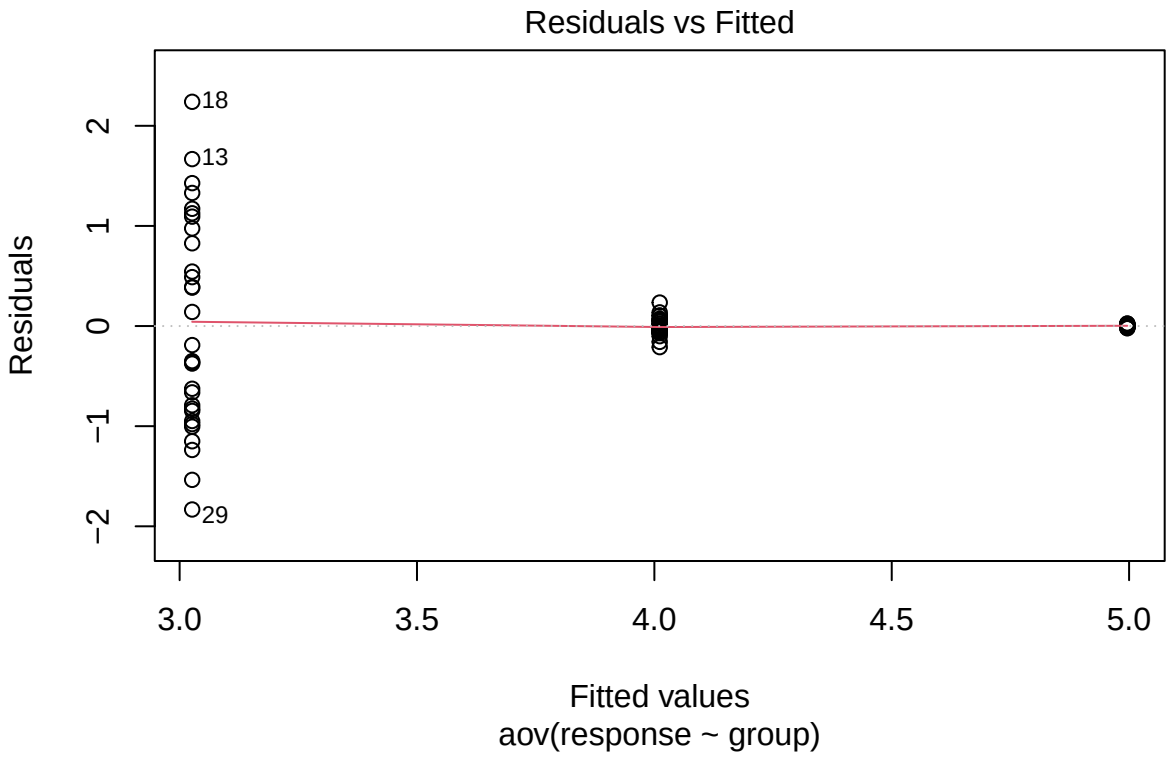
```
##   group response
## 1     1 3.852954
## 2     1 3.167466
## 3     1 2.400217
## 4     1 4.001216
## 5     1 4.119065
## 6     1 3.570592
```

What we want to see is what the diagnostic plots look like if we do an ANOVA. So we start by creating the model:

```
model <- aov(response ~group, data)
```

If we are interested in equality of variance, we want to look at the "Residuals vs Fitted" plot. Recall that we want all the "columns" to be of approximately the same height. Here, we have intentionally created different variances, so we expect this to be reflected in the "Residuals vs Fitted" plot. And indeed:

```
plot(model, 1)
```



For the two bonus questions, the normal-QQ plot is maybe the most difficult to understand. Remind yourself of the concept of quantile: the x quantile of a dataset is the value such that x fraction of your dataset is below that value. For instance, the 0.5 quantile is the median. The QQ plot asks the question: For each of the residuals in your dataset, what quantile are they in? This is the y axis, labelled "sample quantiles". Now, for each of the residuals, if they were truly normally distributed, what quantile *would they be* in? This is the x axis, labelled "theoretical quantiles". You can compute this theoretical quantile distribution by generating a normally distributed sample in R.

Compared to this, the "residual vs fitted" plot is relatively easy to understand. You can take each group mean as the fitted value (x axis) and for each data point, just plot the residual on the y axis.

Interactions

You have been asked to sketch different scenarios. How do you know you are right? One option would be to simulate datasets for all the different scenarios, plot them, and see what they look like. This will help build your intuition.

Last update by DJ MacGregor in 2024