Lab 15 - Factorial ANOVA

Nick Sumpter

2022-09-23

- Today's Lab
- Loading Packages and Data
    - Getting to Know Your Data
- Assumptions
- Running a factorial ANOVA
    - Reporting the Results
- Independent Practice

Today's Lab

In this lab, we will be continuing our series of labs on ANOVA's by introducing the independent factorial ANOVA (a.k.a. two-way independent ANOVA), which incorporates two or more independent grouping variables and tests for a difference in means between any combinations of these grouping variables. It also tests whether the effect of one grouping variable on the outcome depends on the level of the other grouping variable (i.e. the interaction effect).

Loading Packages and Data

```
library(car)
library(ez)
library(pastecs)
library(tidyverse)

theme_set(theme_bw())

setwd("~/Documents/PhD/Teaching/GRD770/R Labs 2022/Lab 15 - Factorial ANOVA")

load("Factorial_ANOVA.RData")

source("../Lab 6 - Normality and Sample Properties 1/functions.R")
```

Getting to Know Your Data

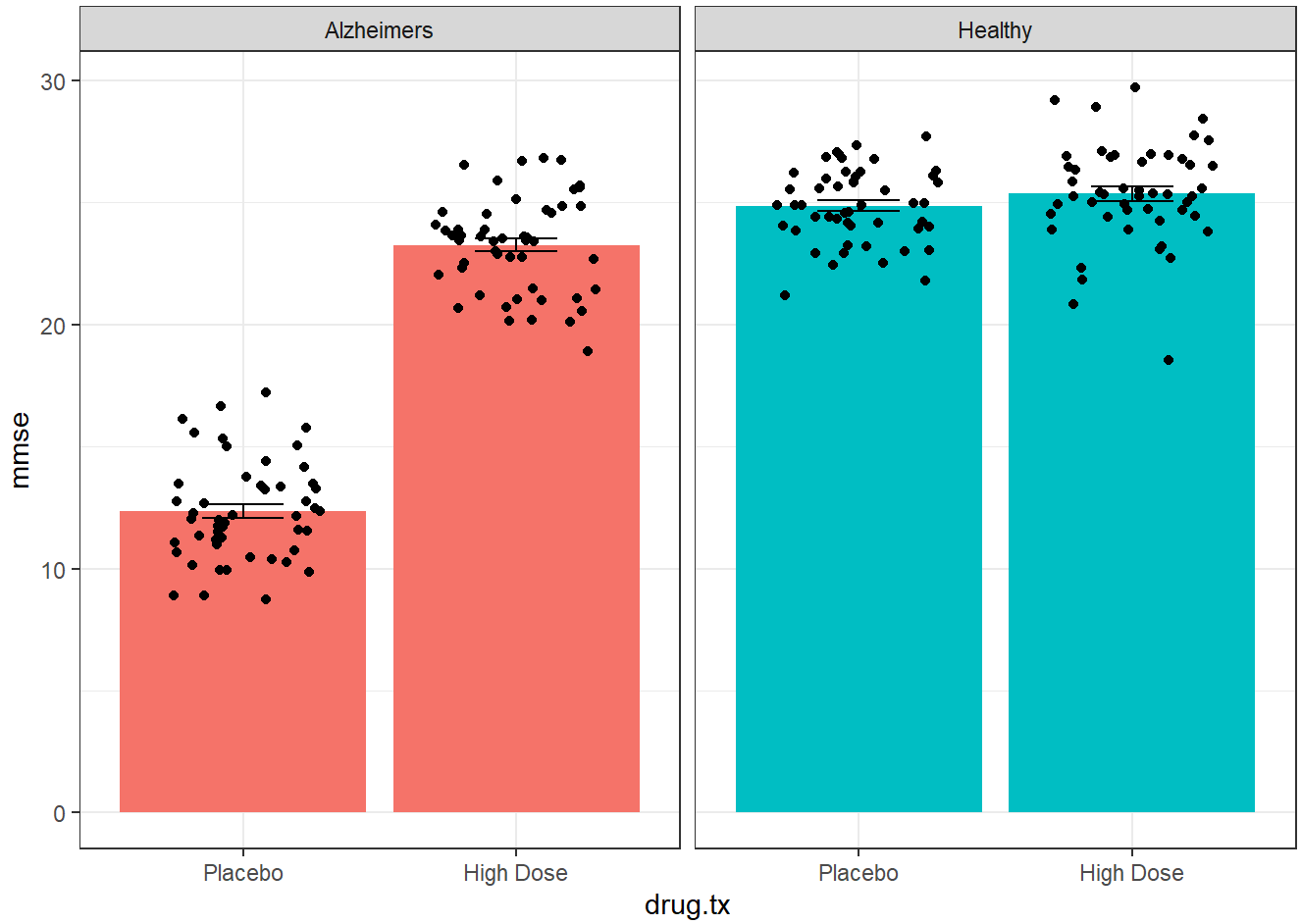
The dataset we will be using is called `ad2`, stored in the `Factorial_ANOVA.RData` file. This dataset contains 4 variables relating to a drug trial that tested the effect of a novel Alzheimer's medication on a cognitive test:

1. `ID` (factor, 200 levels): describes the ID number for all 200 individuals in the study
2. `health` (factor, 2 levels): describes whether the individual has Alzheimer's or is a healthy control
3. `drug.tx` (factor, 2 levels): describes whether the individual was given the trial drug or a placebo

4. `mmse` (numeric): outcome of a cognitive test, measured on a scale of 0-30. 30 means basically no cognitive impairment whereas 0 is extremely severe cognitive impairment

Our hypothesis is that the `mmse` score will be influenced by both `health` status and whether the individual was prescribed the trial drug or placebo (`drug.tx`). We also wish to test whether any effect of the drug on the `mmse` score is dependent on the `health` status. We therefore will plot the group means of `mmse` at all combinations of `health` and `drug.tx`.

```
ggplot(data = ad2, mapping = aes(x = drug.tx, y = mmse)) +
  geom_bar(mapping = aes(fill = health), stat = "summary", fun = "mean", show.legend = F
ALSE) +
  geom_errorbar(stat = "summary", fun.data = "mean_se", width = 0.3) +
  geom_jitter(width = 0.3) +
  facet_wrap(~ health)
```



As you can see, there seems to be a clear difference in means between drug doses for the Alzheimer's group, but this difference seems to almost completely disappear among the healthy controls. Let's see what our model tells us.

Assumptions

1. Independence (check study design - met in this case as each row refers to a different individual in the dataset)

2. Normality of the model residuals (tested after running the model)
3. Homogeneity of variance (tested after running the model)

Running a factorial ANOVA

This is exactly the same as for the one-way independent ANOVA but instead of one variable for the `between` argument, you provide two variables wrapped in the `.` function.

```
mod <- ezANOVA(data = ad2,
               dv = mmse,
               between = .(health, drug.tx),
               wid = ID,
               type = 3,
               return_aov = TRUE)

mod
```

```
## $ANOVA
##           Effect DFn DFd        F            p p<.05       ges
## 2         health   1 196 733.7465 3.527397e-68     * 0.7891898
## 3        drug.tx   1 196 447.8651 1.630518e-52     * 0.6955884
## 4 health:drug.tx   1 196 373.0069 3.057674e-47     * 0.6555402
##
## $`Levene's Test for Homogeneity of Variance`
##   DFn DFd      SSn      SSd         F         p p<.05
## 1   3 196 3.989632 286.1227 0.9109937 0.4366698
##
## $aov
## Call:
##    aov(formula = formula(aov_formula), data = data)
##
## Terms:
##                     health   drug.tx health:drug.tx Residuals
## Sum of Squares  2665.3131 1626.8571      1354.9369  711.9644
## Deg. of Freedom         1         1              1       196
##
## Residual standard error: 1.905904
## Estimated effects may be unbalanced
```

From this output, we can immediately see that there is a significant main effect of the `health` variable and the `drug.tx` variable on `mmse`. This is shown by the p-value in the first and second row of the output table being less than 0.05. This can be interpreted as a difference in mean `mmse` score when looking across levels of either `drug.tx` or `health`.

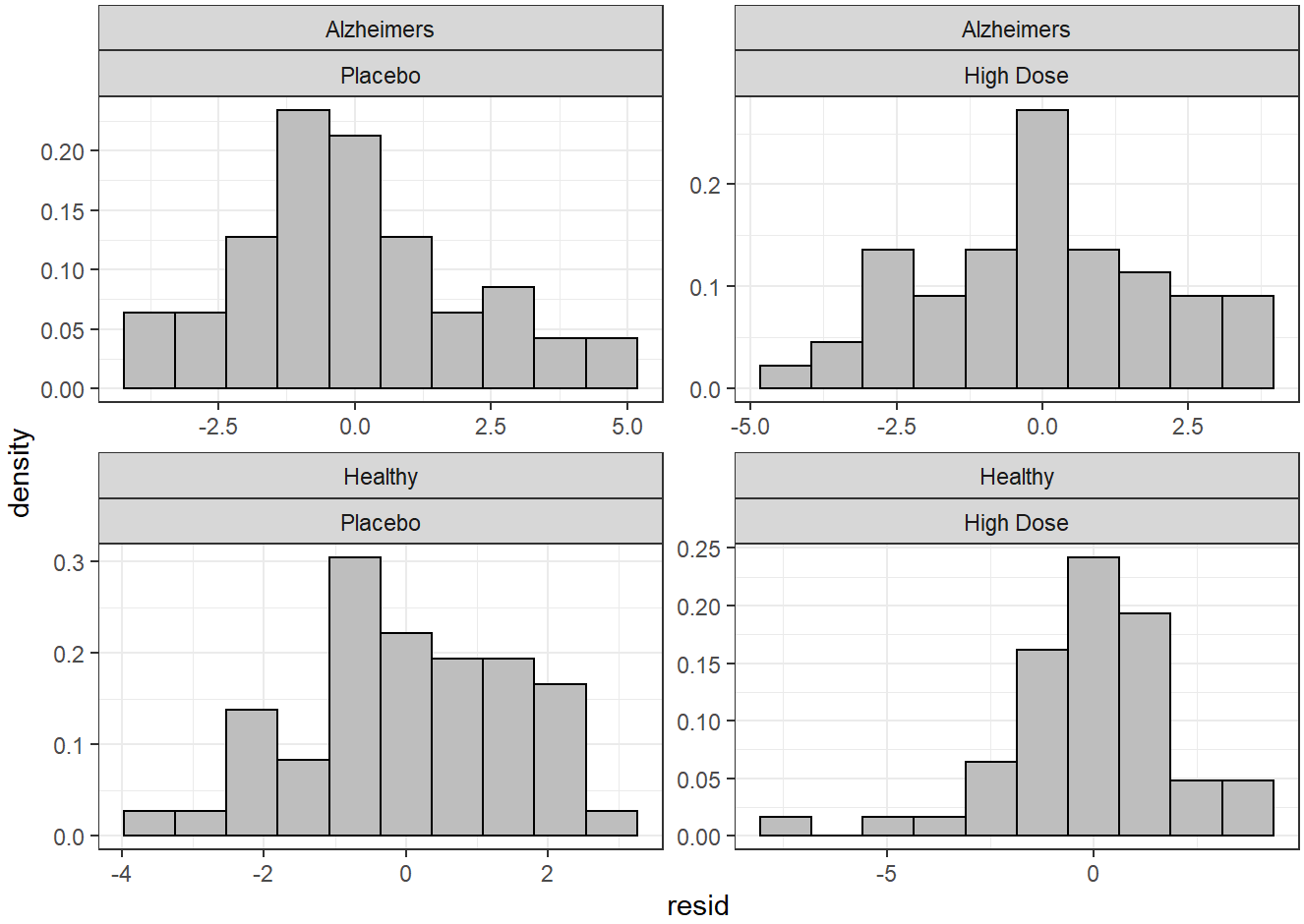
The third row of the ANOVA table shows us the model parameters for the `health:drug.tx` interaction term. As you can see, this also has a highly significant p-value of 3.06e-47. This tells us that the effect of either of these variables on `mmse` depends on the level of the other variable. We can interpret this as the drug having a differential effect on cognitive ability in Alzheimer's patients in comparison to healthy controls. Based on the plot

we made earlier, this makes sense, and we can see that the effect seems to be stronger in Alzheimer's patients compared to healthy controls. Of course, we would need to run post-hoc tests to confirm this (which we will do in two labs time).

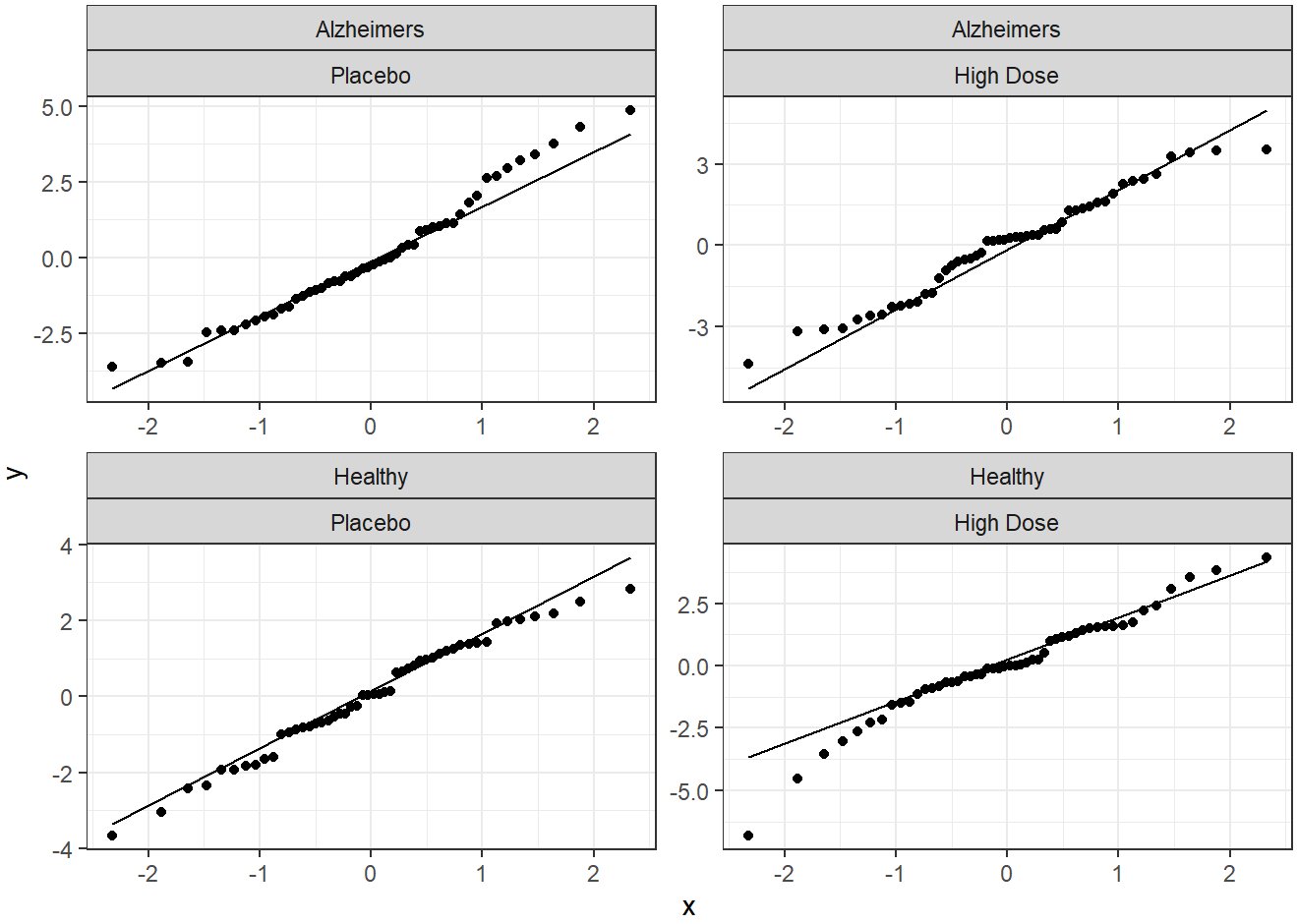
Importantly, as the interaction effect was significant, this essentially trumps the significant main effects. Given that the effect of these variables depends on the level of the other, it now no longer makes sense to interpret the main effect of either of these variables on their own. We will still be reporting these main effects regardless.

Note that the Levene's test was non-significant and thus the homogeneity of variance assumption was met. We finally need to test the normality assumption of the residuals:

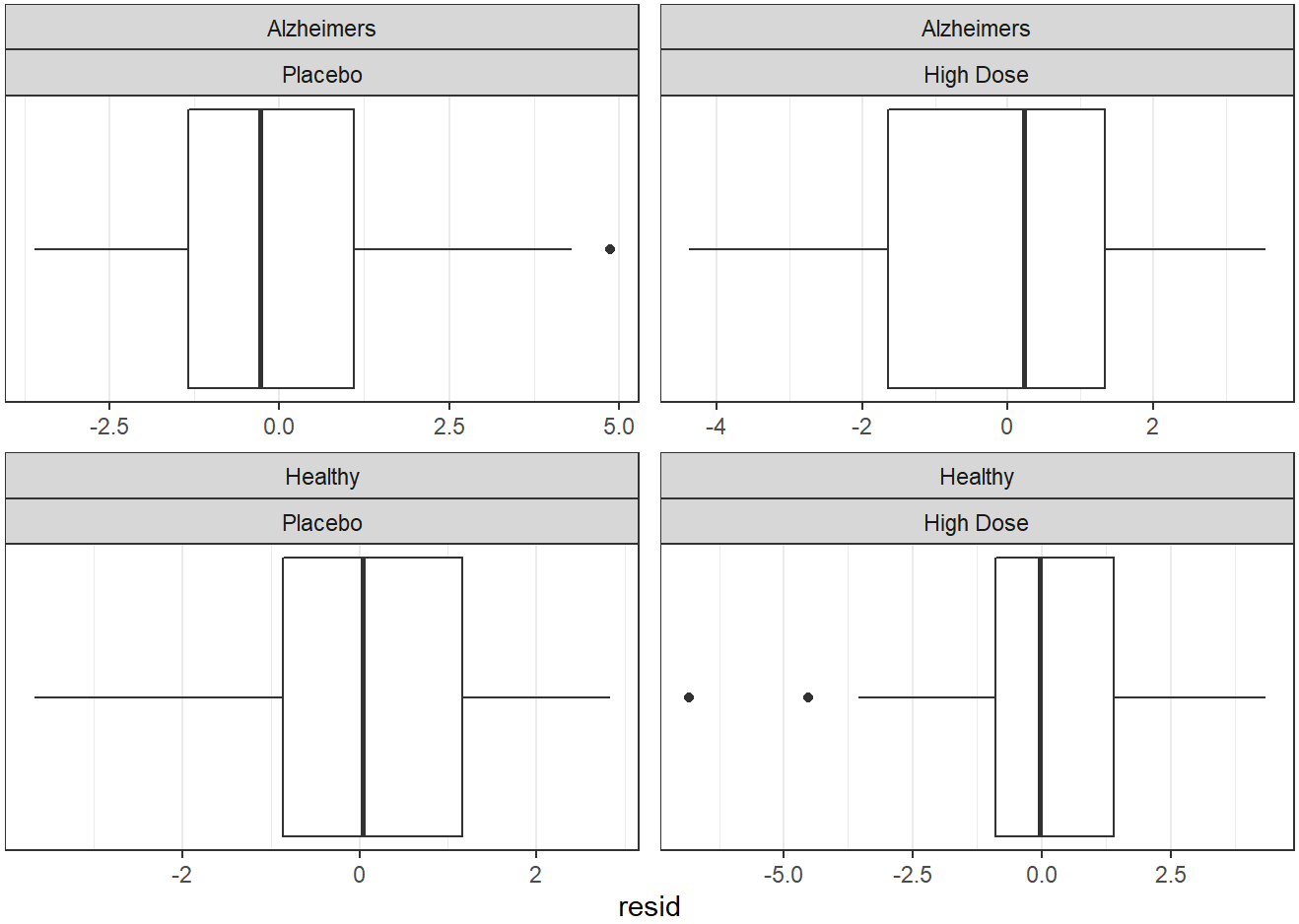
```
# Extracting residuals column
residuals <- tibble(group1 = ad2$health,
                    group2 = ad2$drug.tx,
                    resid = resid(mod$aov))

# Normality within groups
ggplot(data = residuals, mapping = aes(x = resid)) +
  geom_histogram(mapping = aes(y = ..density..), bins = 10, fill = 'gray', color = 'blac
k') +
  facet_wrap(~ group1 + group2, scales = "free")
```



```
ggplot(data = residuals, mapping = aes(sample = resid)) +
  geom_qq() +
  geom_qq_line() +
  facet_wrap(~ group1 + group2, scales = "free")
```




```
ggplot(data = residuals, mapping = aes(x = resid)) +
  geom_boxplot() +
  theme(axis.ticks.y = element_blank(),
        axis.text.y = element_blank(),
        panel.grid.major.y = element_blank(),
        panel.grid.minor.y = element_blank()) +
  facet_wrap(~ group1 + group2, scales = "free")
```




```
stat.desc.clean(dataset = residuals, variable = resid, group1, group2)
```

```
## # A tibble: 4 × 8
## # Groups:   group1, group2 [4]
##   group1     group2    skewness skew.2SE kurtosis kurt.2SE normtest.W normtest.p
##   <fct>      <ord>        <dbl>    <dbl>    <dbl>    <dbl>      <dbl>      <dbl>
## 1 Alzheimers Placebo      0.426    0.632   -0.452   -0.342      0.973      0.305
## 2 Alzheimers High Dose   -0.110   -0.163   -0.753   -0.569      0.974      0.347
## 3 Healthy    Placebo     -0.259   -0.385   -0.629   -0.475      0.981      0.614
## 4 Healthy    High Dose   -0.609   -0.904    1.29     0.972      0.962      0.109
```

Reporting the Results

“There was a significant main effect of drug treatment on mmse score, F(1,196) = 447.9, p = 1.6e-52. The post-hoc tests revealed… ”

“There was a significant main effect of health status (Alzheimer’s vs healthy control) on mmse score, F(1,196) = 733.7, p = 3.5e-68. The post-hoc tests revealed… ”

“There was a significant interaction effect of health status and drug treatment on mmse score, F(1,196) = 373.0, p = 3.1e-47. This indicates that the effect of drug treatment on mmse score depended on the health status of the individual. ”

Independent Practice

For your independent practice, we want you to perform a two-way independent factorial ANOVA in the `jobsatisfaction` dataset. We want you to test the effect of `gender` and `education_level` on job satisfaction `score`, and whether the effect of either of these variables depends on the level of the other.

1. Get to know your data: describe your variables and their distribution
2. Run the factorial ANOVA
3. Assumptions
    a. Determine whether assumptions are met
    b. If assumptions are not met, describe what you will do to account for this. If possible, modify your model to meet the assumptions
4. Report your findings as you would describe them in the results section