Lab 12 - ANOVA

Nick Sumpter (Edited by Eddie-Williams Owiredu & Guy Twa)

2023-10-02

- Today's Lab
- Loading Packages and Data
    - Getting to Know Your Data
- Assumptions for one-way independent ANOVA
- Running a one-way independent ANOVA
    - Assessing Normality
    - Reporting an ANOVA
- Independent Practice

Today's Lab

Today we will learn how to use Analysis of Variance (ANOVA) to analyze our data. ANOVA models test whether there is a significant difference in the means between two or more groups, which is a generalization of the t-test. They can also handle tests across multiple grouping variables, both within-subjects and between-subjects, as well as account for influences of continuous variables on the outcome measure. In today's lab, we will show you how to perform a basic one-way ANOVA using the `ezANOVA` function from the `ez` package. We will cover the basic example of testing a difference in means among 3 between-subjects groups.

Loading Packages and Data

For this lab, we will be using four packages (car, ez, pastecs, and tidyverse). The ez package is new and so you will need to install it:

```
install.packages("ez")
```

The dataset we will be using in the example portion of the lab is the `mtcars` dataset, which is built into R. For your independent practice, you will need to download the `ANOVA_diet.RData` dataset from Canvas.

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

theme_set(theme_bw())

setwd("/Users/eddie-williamsowiredu/Desktop/grd770_23/Lab12")

source("functions.R")

mtcars <- mtcars
```

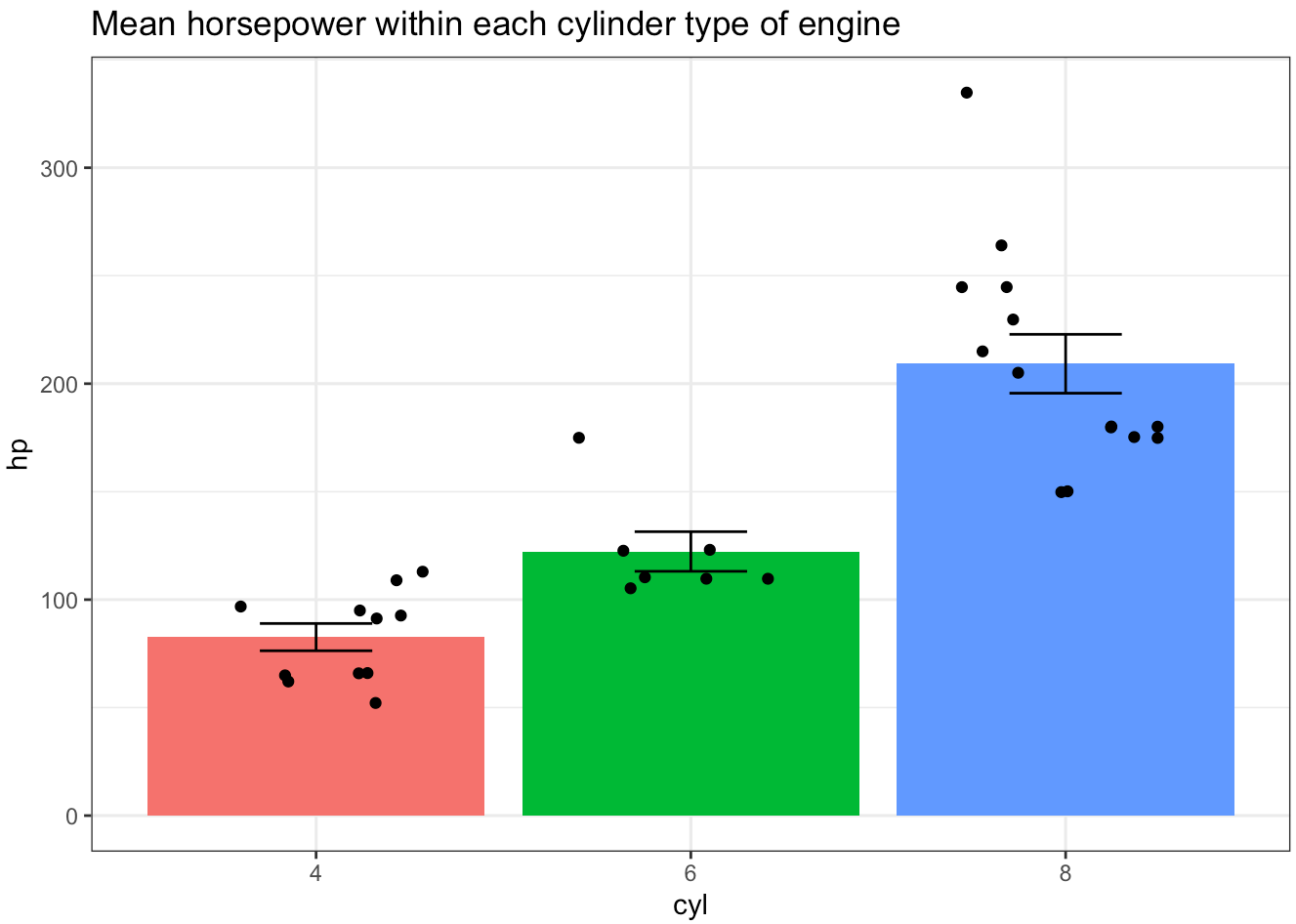
Getting to Know Your Data

Though we have come across this dataset before, we will look at the summary of the full dataset, then plot the mean horsepower (`hp`) within each cylinder group (`cyl`). We'll need to modify the cyl variable to be a factor prior to running our analyses.

```
summary(mtcars)
```

```
##       mpg             cyl             disp             hp       
##  Min.   :10.40   Min.   :4.000   Min.   : 71.1   Min.   : 52.0  
##  1st Qu.:15.43   1st Qu.:4.000   1st Qu.:120.8   1st Qu.: 96.5  
##  Median :19.20   Median :6.000   Median :196.3   Median :123.0  
##  Mean   :20.09   Mean   :6.188   Mean   :230.7   Mean   :146.7  
##  3rd Qu.:22.80   3rd Qu.:8.000   3rd Qu.:326.0   3rd Qu.:180.0  
##  Max.   :33.90   Max.   :8.000   Max.   :472.0   Max.   :335.0  
##       drat             wt             qsec             vs        
##  Min.   :2.760   Min.   :1.513   Min.   :14.50   Min.   :0.0000  
##  1st Qu.:3.080   1st Qu.:2.581   1st Qu.:16.89   1st Qu.:0.0000  
##  Median :3.695   Median :3.325   Median :17.71   Median :0.0000  
##  Mean   :3.597   Mean   :3.217   Mean   :17.85   Mean   :0.4375  
##  3rd Qu.:3.920   3rd Qu.:3.610   3rd Qu.:18.90   3rd Qu.:1.0000  
##  Max.   :4.930   Max.   :5.424   Max.   :22.90   Max.   :1.0000  
##        am              gear            carb      
##  Min.   :0.0000   Min.   :3.000   Min.   :1.000  
##  1st Qu.:0.0000   1st Qu.:3.000   1st Qu.:2.000  
##  Median :0.0000   Median :4.000   Median :2.000  
##  Mean   :0.4062   Mean   :3.688   Mean   :2.812  
##  3rd Qu.:1.0000   3rd Qu.:4.000   3rd Qu.:4.000  
##  Max.   :1.0000   Max.   :5.000   Max.   :8.000
```

```
mtcars2 <- mtcars %>%
  mutate(cyl = factor(cyl))

ggplot(data = mtcars2, mapping = aes(x = cyl, y = hp)) +
  geom_bar(mapping = aes(fill = cyl), stat = "summary", fun = "mean", show.legend = FALS
E) +
  geom_errorbar(stat = "summary", fun.data = "mean_se", width = 0.3) +
  geom_jitter(width = 0.3) +
  labs(title = "Mean horsepower within each cylinder type of engine") +
  theme_bw()
```



Assumptions for one-way independent ANOVA

First, let's go over the assumptions that must be met in order to perform an ANOVA. A standard between-subjects one-way ANOVA will have the following 3 assumptions:

1. Independence
2. Normality of model residuals within groups
3. Homogeneity of variance of model residuals across groups

Each of these assumptions has been covered previously. You can determine independence from the design of the experiment. Normality uses the same graphical and mathematical tests we have in previous labs, making sure to test the model residuals at each level of the grouping variable. Finally, the homogeneity of variance assumption requires the use of Levene's test.

Note: If the normality assumption is violated in one of the groups and you want to fix this using a mathematical transformation like a square root or log then you need to apply that same transformation to every group and retest normality in each group. You cannot take the square root of the values in a single group and then compare the mean of the square root values to the raw values of other groups.

Running a one-way independent ANOVA

We will test to see if there is a difference in mean horsepower based on engine type (4, 6, or 8 cylinder engines) The `ez` function can be used in several different ways, but we will input the following arguments:

- `data` : the dataset to use
- `dv` : the name of the dependent variable
- `between` : the name of the between-subject variable
- `wid` : ID variable, will need to make this if your dataset is missing one
- `type` : type of sums of squares to calculate, make sure to set to '3'
- `return_aov` : whether we want the model to calculate another type of R object called an `aov` , set to TRUE

```
mtcars3 <- mtcars2 %>%
  mutate(ID = factor(1:nrow(.)))

mod <- ezANOVA(data = mtcars3,
               dv = hp,
               between = cyl,
               wid = ID,
               type = 3,
               return_aov = TRUE)

mod
```

```
## $ANOVA
##   Effect DFn DFd        F            p p<.05       ges
## 2    cyl   2  29 36.17687 1.318541e-08     * 0.7138734
##
## $`Levene's Test for Homogeneity of Variance`
##   DFn DFd      SSn      SSd        F          p p<.05
## 1   2  29 4235.047 21095.92 2.910903 0.07045752
##
## $aov
## Call:
##    aov(formula = formula(aov_formula), data = data)
##
## Terms:
##                       cyl Residuals
## Sum of Squares  104030.54  41696.33
## Deg. of Freedom         2        29
##
## Residual standard error: 37.91839
## Estimated effects may be unbalanced
```

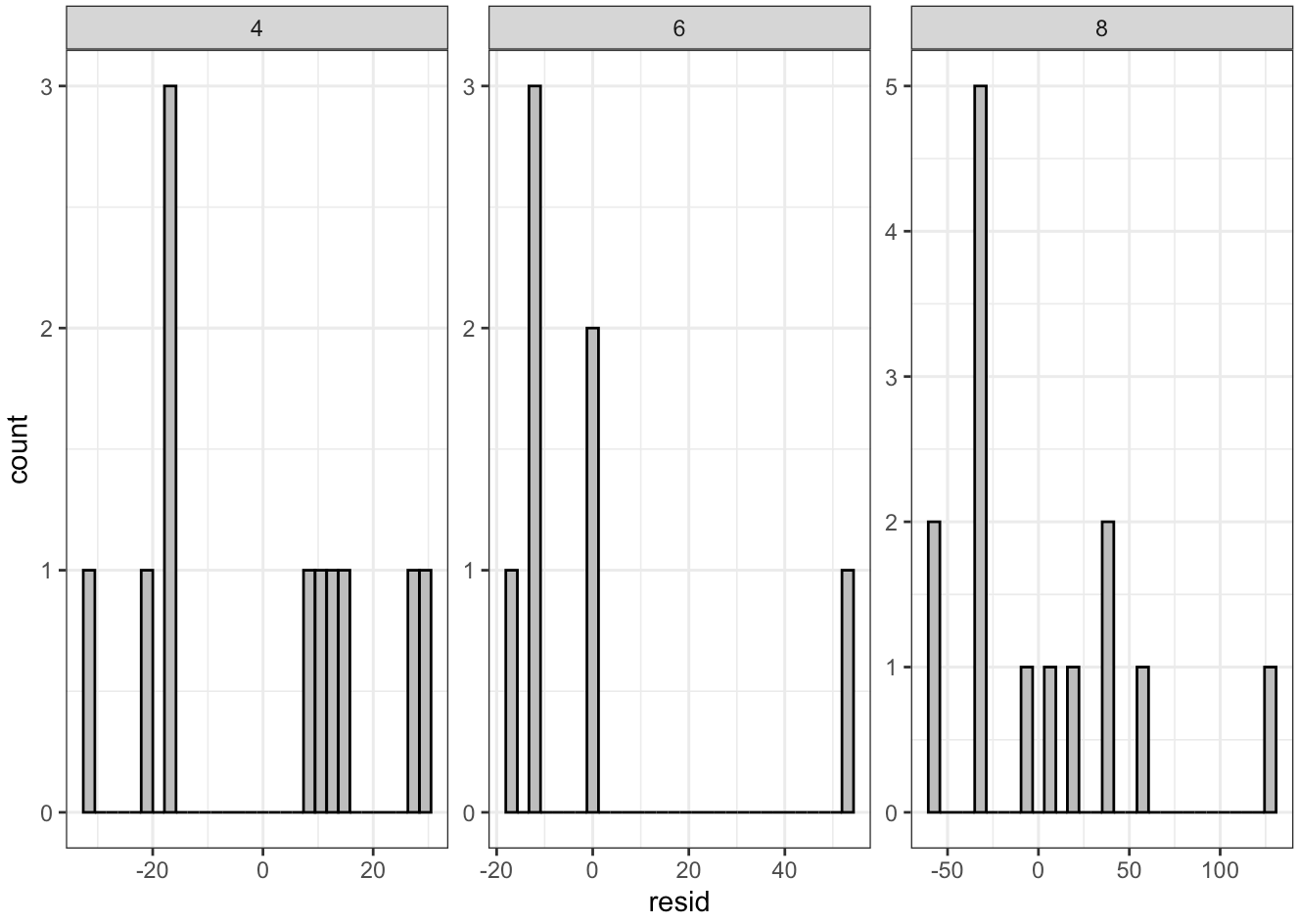
You will see 3 separate tables in the output.

1. The first (`ANOVA`) includes degrees of freedom, the F-statistic, and p value for the main effect of cyl on hp. If the main effect is significant, it does not tell us explicitly which groups are different, only that there is an overall difference among the groups. In order to see where the differences are, post-hoc tests are run, but that is covered later in the semester.
2. The second (`Levene's Test for Homogeneity of Variance`) shows the Levene's test output. In this example, the homogeneity of variance assumption is met based on the non-significant Levene's test.
3. The third (`aov`) will be used to extract residuals for testing normality.

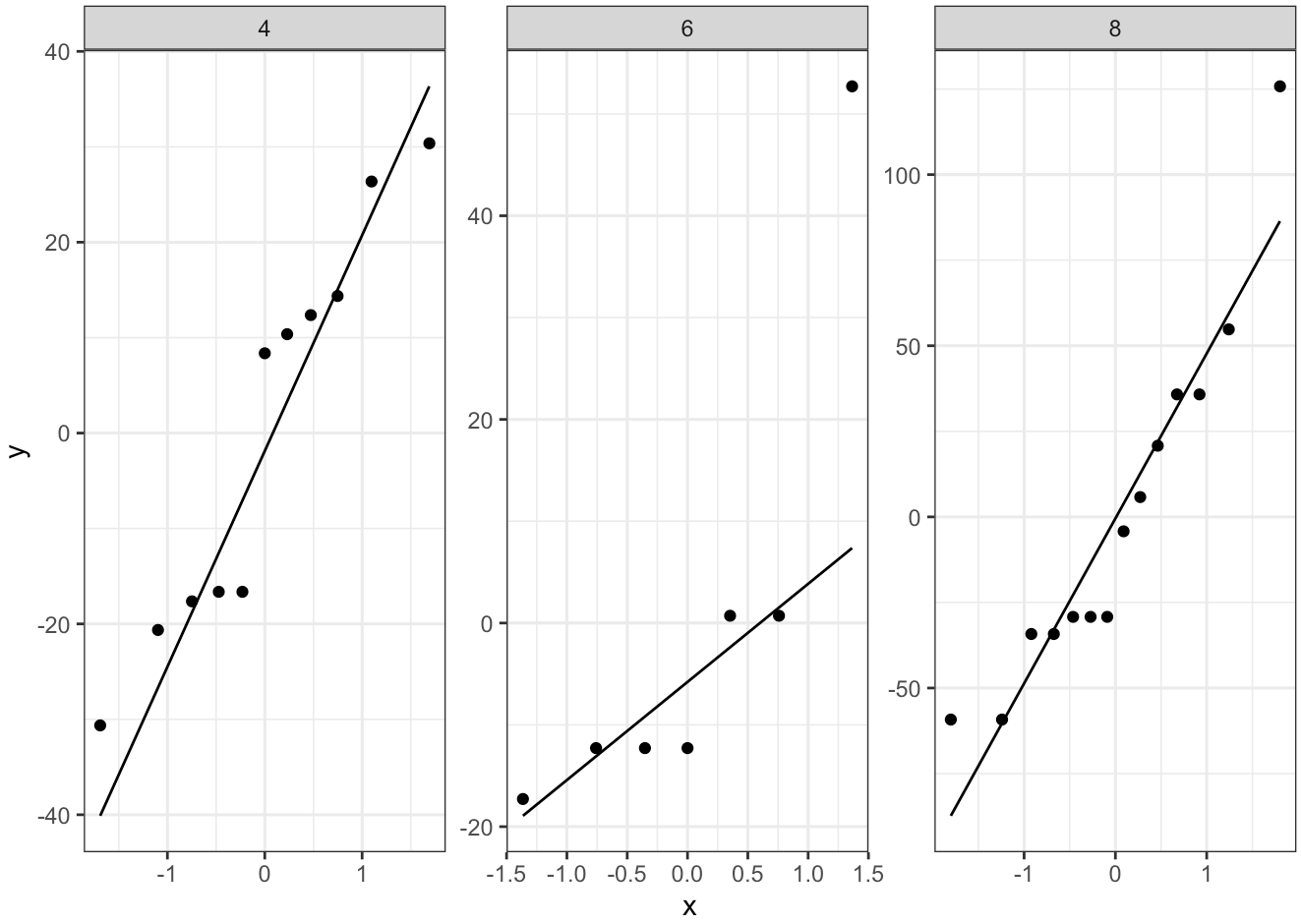
Assessing Normality

We can obtain the model residuals easily by using the `resid` function on the `aov` object that ezANOVA outputs. To access this `aov` object we simply use the `$` function (`mod$aov`). We'll make a small table with one column for the grouping variable and one column for the residuals.

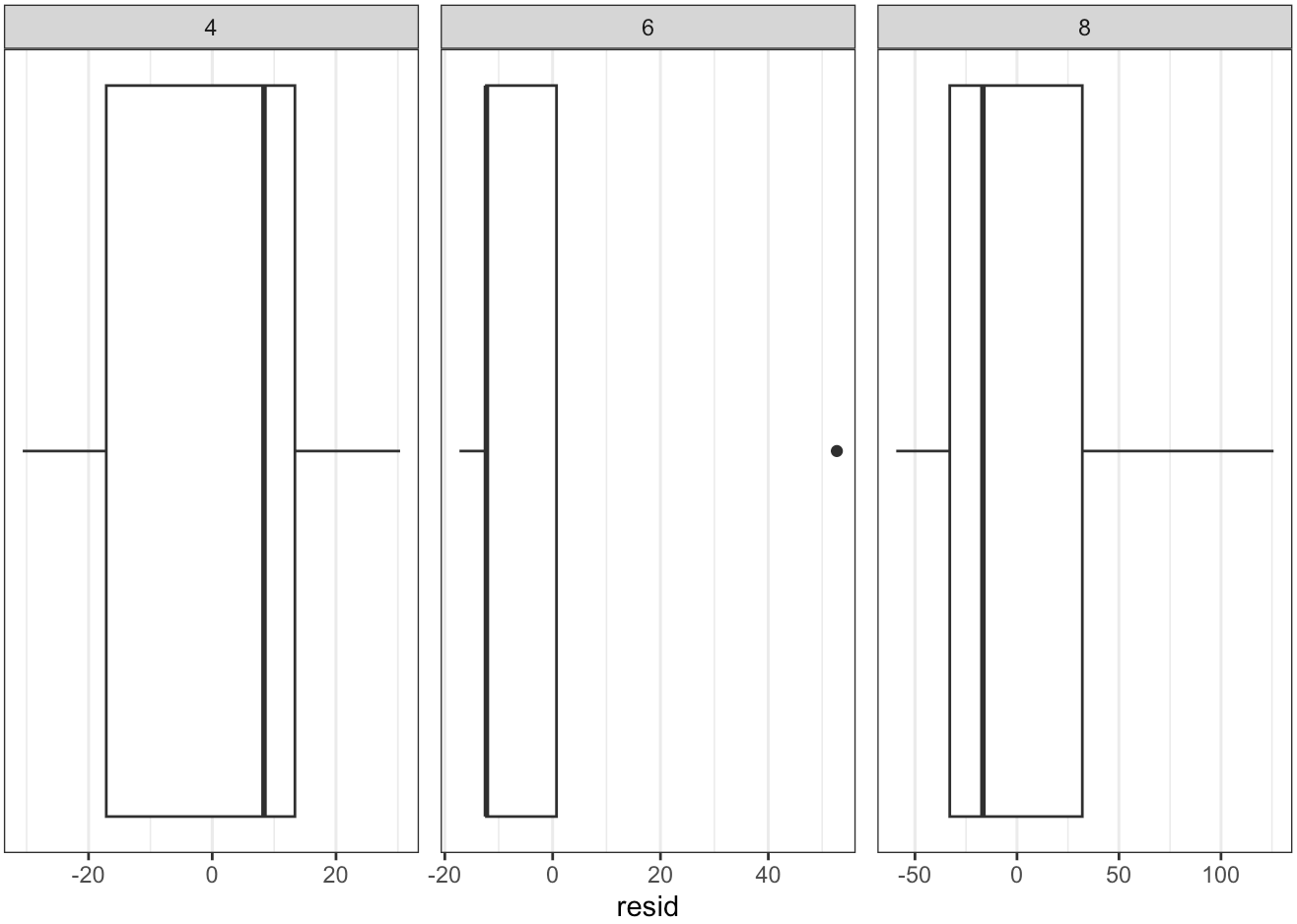
```
# Extracting residuals
residuals <- tibble(group = mtcars2$cyl,
                    resid = resid(mod$aov))

# Testing Normality
ggplot(data = residuals, mapping = aes(x = resid)) +
  geom_histogram(bins = 30, fill = 'gray', color = 'black') +
  facet_wrap(~ group, scales = "free")
```



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




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

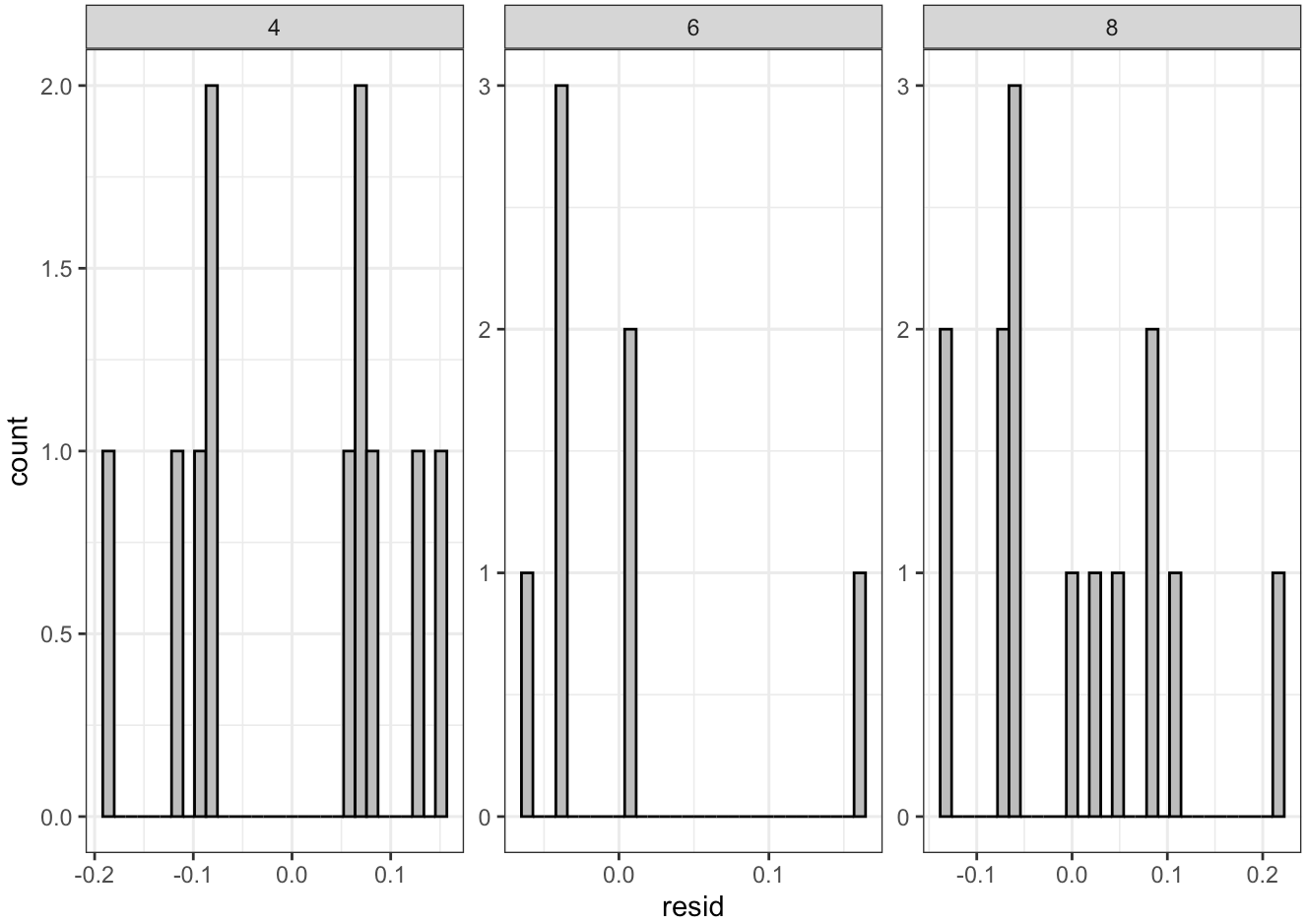
```
## # A tibble: 3 × 7
## # Groups:   group [3]
##   group skewness skew.2SE kurtosis kurt.2SE normtest.W normtest.p
##   <fct>    <dbl>    <dbl>    <dbl>    <dbl>      <dbl>      <dbl>
## 1 4      0.00626  0.00474  -1.71    -0.668       0.906    0.219
## 2 6      1.36     0.856     0.249    0.0785      0.691    0.00294
## 3 8      0.909    0.761     0.0921   0.0399      0.898    0.105
```

The normality assumption seems slightly off in the 6 cylinder group, but this seems to be exclusively due to one outlier and so it is probably okay to say this assumption is met. However, let us go ahead with log normalization of the dependent variable to see if we are able to get around the non-parametric nature of cyl 6.

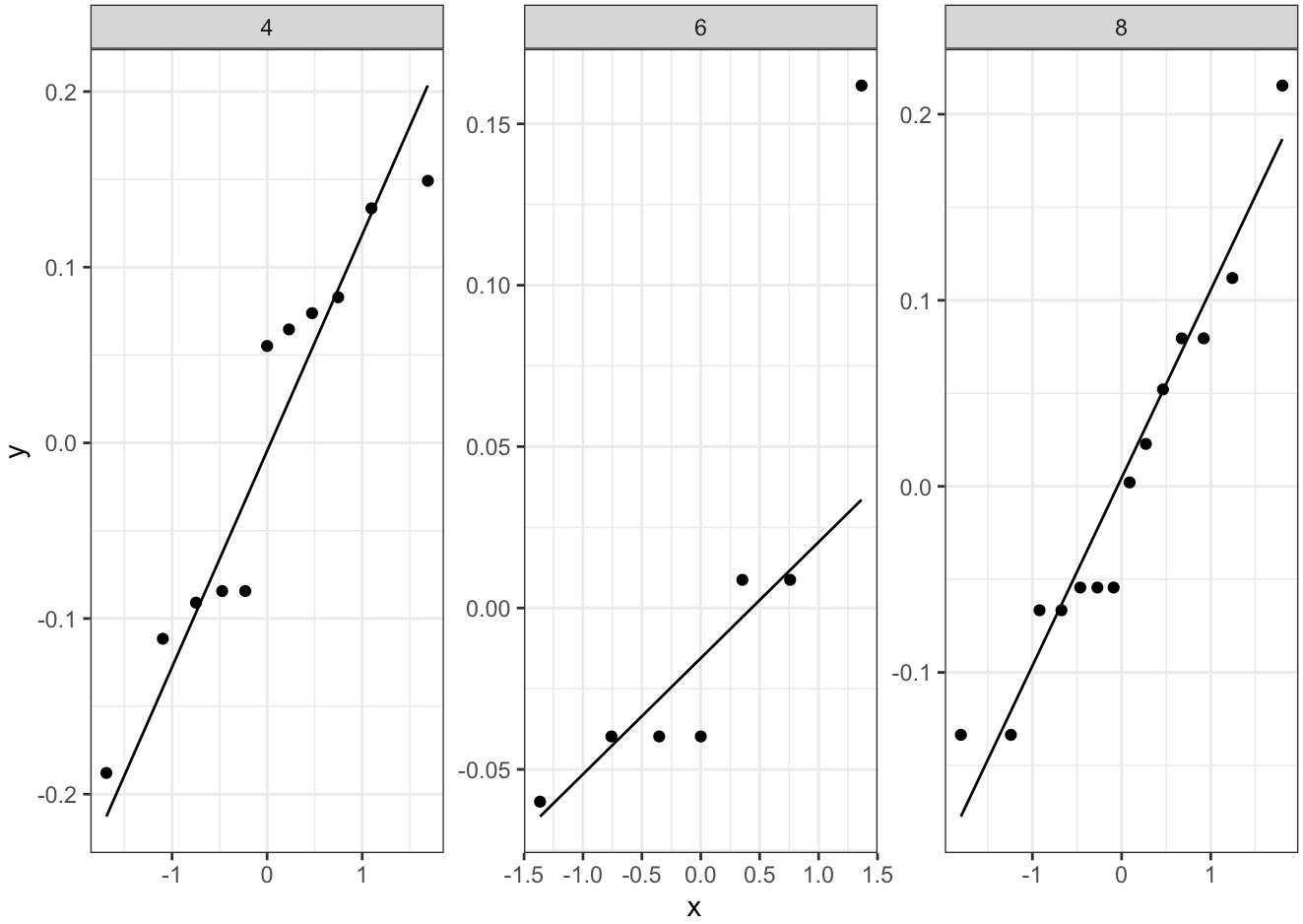
```r
# Adding the log10 of the hp variable
mtcars4 <- mtcars3 %>% mutate(log_hp = log10(hp))

mod1 <- ezANOVA(data = mtcars4,
                dv = log_hp,
                between = cyl,
                wid = ID,
                type = 3,
                return_aov = TRUE)

mod1
```

```
## $ANOVA
##   Effect DFn DFd        F            p p<.05       ges
## 2    cyl   2  29 50.89638 3.268188e-10     * 0.7782752
##
## $`Levene's Test for Homogeneity of Variance`
##   DFn DFd       SSn       SSd        F        p p<.05
## 1   2  29 0.0112747 0.1389084 1.176913 0.322521
##
## $aov
## Call:
##    aov(formula = formula(aov_formula), data = data)
##
## Terms:
##                       cyl Residuals
## Sum of Squares  1.0277266 0.2927917
## Deg. of Freedom         2        29
##
## Residual standard error: 0.1004802
## Estimated effects may be unbalanced
```

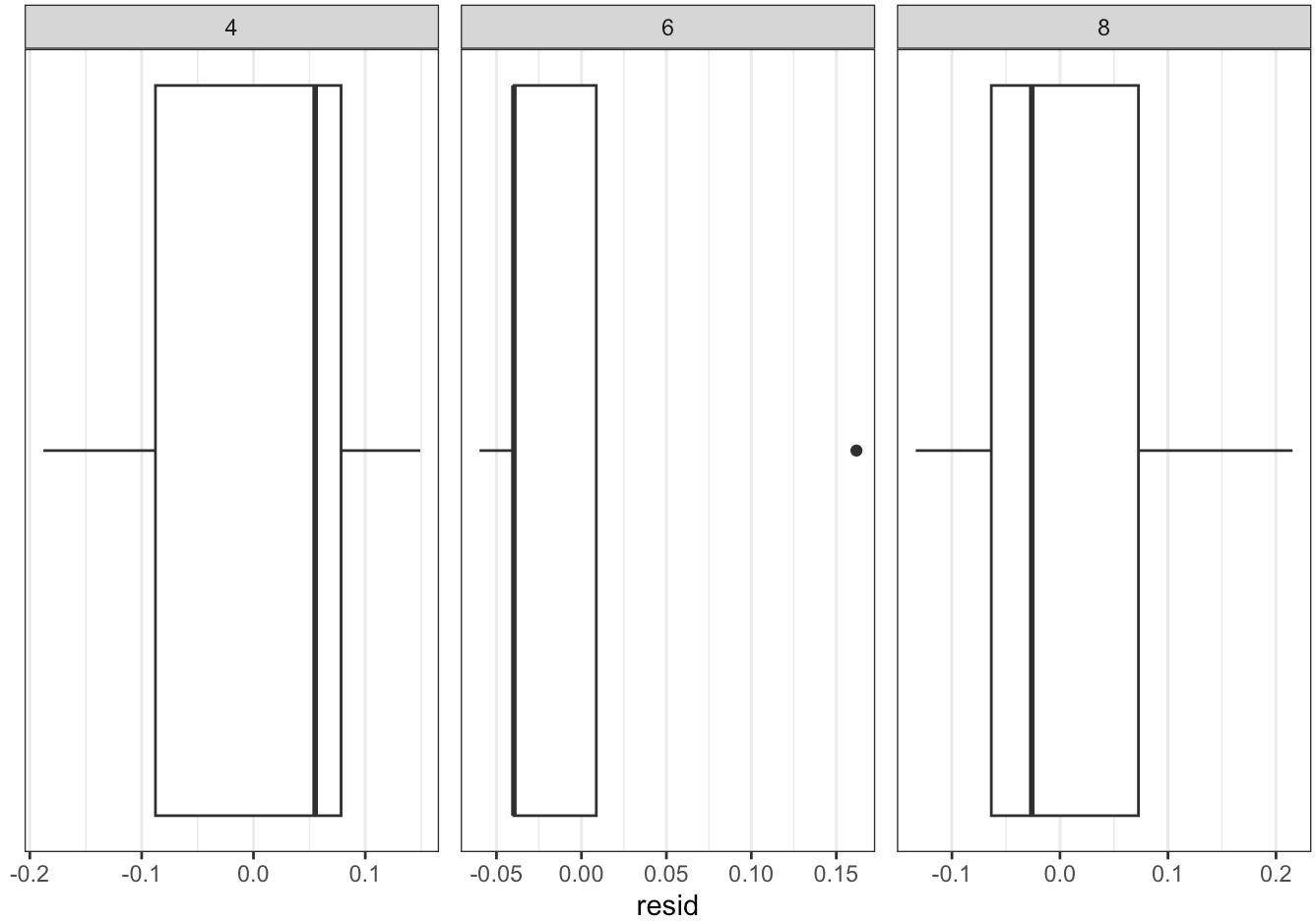
```r
# Extracting residuals
residuals <- tibble(group = mtcars2$cyl,
                    resid = resid(mod1$aov))

# Testing Normality
ggplot(data = residuals, mapping = aes(x = resid)) +
  geom_histogram(bins = 30, fill = 'gray', color = 'black') +
  facet_wrap(~ group, scales = "free")
```




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




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

```
## # A tibble: 3 × 7
## # Groups:   group [3]
##   group skewness skew.2SE kurtosis kurt.2SE normtest.W normtest.p
##   <fct>    <dbl>    <dbl>    <dbl>    <dbl>      <dbl>      <dbl>
## 1 4       -0.176   -0.133  -1.66   -0.648        0.903    0.198
## 2 6        1.25     0.785   0.0196  0.00619      0.736    0.00899
## 3 8        0.490    0.410  -0.725  -0.314        0.940    0.416
```

We can see here that log-normalization did not have much effect on the distribution of the cyl 6. Try your hands on other forms of normalization and see what you get.

Reporting an ANOVA

First let's get the mean horsepower for each cylinder type in a similar way to last lab.

```
mtcars2 %>%
  group_by(cyl) %>%
  summarize(n = n(),
            mean = mean(hp),
            sd = sd(hp))
```

```
## # A tibble: 3 × 4
##   cyl       n  mean    sd
##   <fct> <int> <dbl> <dbl>
## 1 4        11  82.6  20.9
## 2 6         7 122.   24.3
## 3 8        14 209.   51.0
```

If there is a significant effect according to the ANOVA output you can report it with the following form:

“There was a significant main effect of <independent variable> on <output variable>, F(<DFn>,<DFd>) = <F-stat>, p = <p-val>.”

DFn is the degrees of freedom for the grouping variable while DFd is the degrees of freedom in the residuals. Typically, there is also more to add after performing post-hoc tests, but those are covered in a later lab.

For this example, we would say: “The mean horsepower (± sd) for each cylinder type was 82.6 ± 20.9 for 4 cylinders, 122.3 ± 24.3 for 6 cylinders, and 209.2 ± 51.0 for 8 cylinder engines. There was a significant main effect of engine type on horsepower, F(2,29) = 36.18, p = 1.32e-08.”

Independent Practice

For these exercises, we are going to be using the `ANOVA-diet` dataset, which describes mouse liver weight after being fed a diet `low` in carbohydrates, `high` in carbohydrates, or with standard mouse food containing a medium level of carbohydrates (`ctrl`). There are 15 individual mice in each group (total of 45 mice). Our hypothesis for these data is that carbohydrate content will cause a significant difference in liver weight. We want to see if mean liver weight significantly differs based on whether a diet is low or high in carbohydrate content. Investigate this using a standard one-way ANOVA.

1. Get to know your data: describe your variables and their distribution
2. Run the ANOVA
3. Assumptions
    a. Determine whether the assumptions of this test 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