Lab 13 - ANCOVA

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

2023-10-04

Today’s Lab

Today we will cover Analysis of Covariance (ANCOVA). ANCOVA is essentially an ANOVA with the addition of a linear covariate in the model. In our example dataset, we will be testing whether the mean body fat of a mouse varies with respect to the environment that it was raised. We will adjust for lean body mass (our linear covariate) in the ANCOVA.

Loading Packages and Data

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

```
install.packages("rstatix")
```

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

theme_set(theme_bw())

setwd("/Users/eddie-williamsowiredu/Desktop/grd770_23/Lab13")

load("ANCOVA_Data.RData")

source("functions.R")
```

Getting to Know Your Data

The data for today is called `mouse.mass` , located in the file “ANCOVA_Data.RData”. This dataset has the following variables in it:

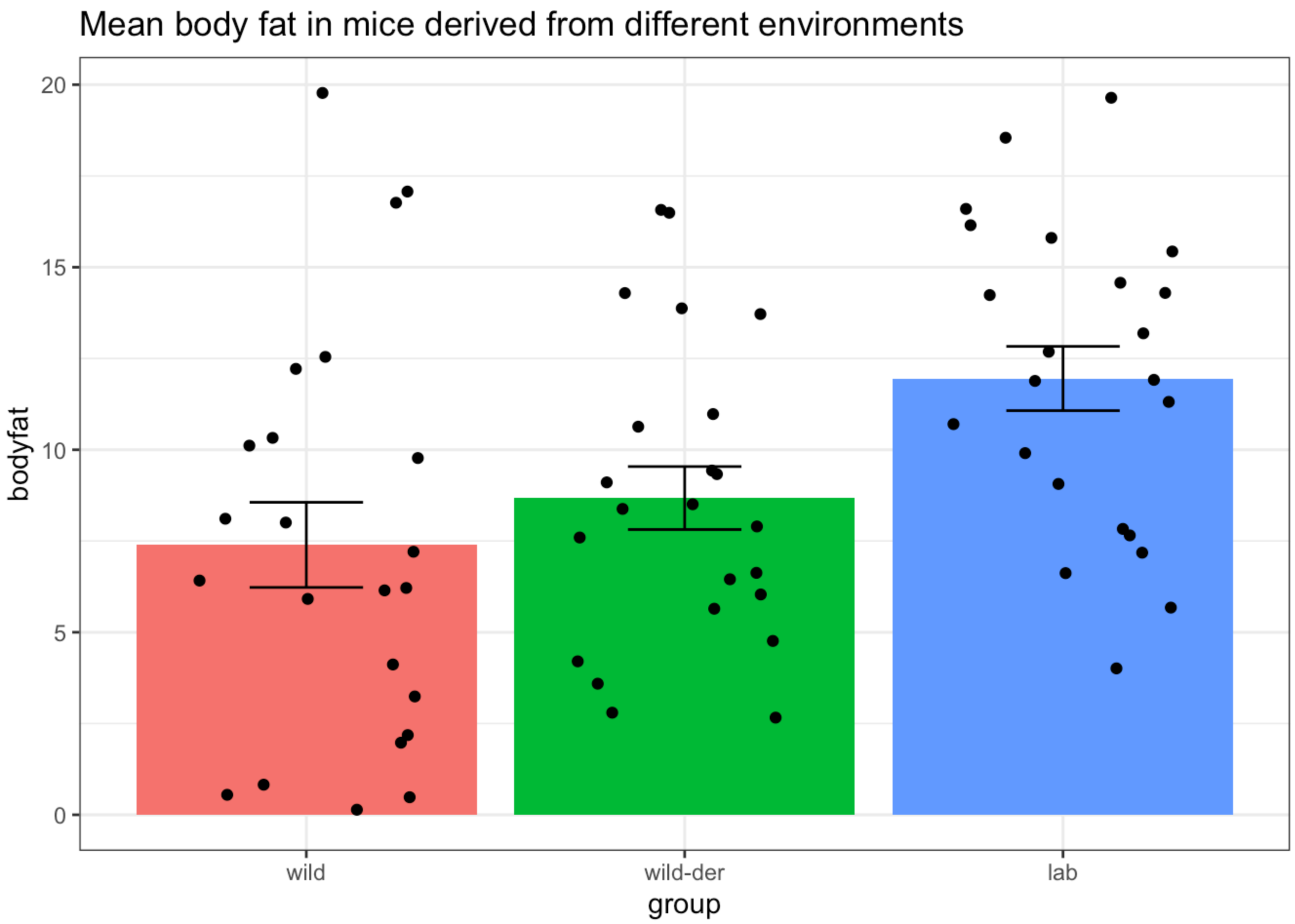
- `idnum` (factor, 69 levels): a unique ID number for each mouse
- `group` (factor, 3 levels): a designation of whether the mouse was fully raised in the lab (‘lab’), in the wild (‘wild’) or was in the wild but has lived in lab for a few weeks (‘wild-der’)
- `bodyfat` (numeric): a measure of the amount of body fat on each mouse
- `leanmass` (numeric): a measure of the mass of the mouse after a period of food deprivation

Let’s have a quick look at the mean value and distribution of the two variables of interest within each group:

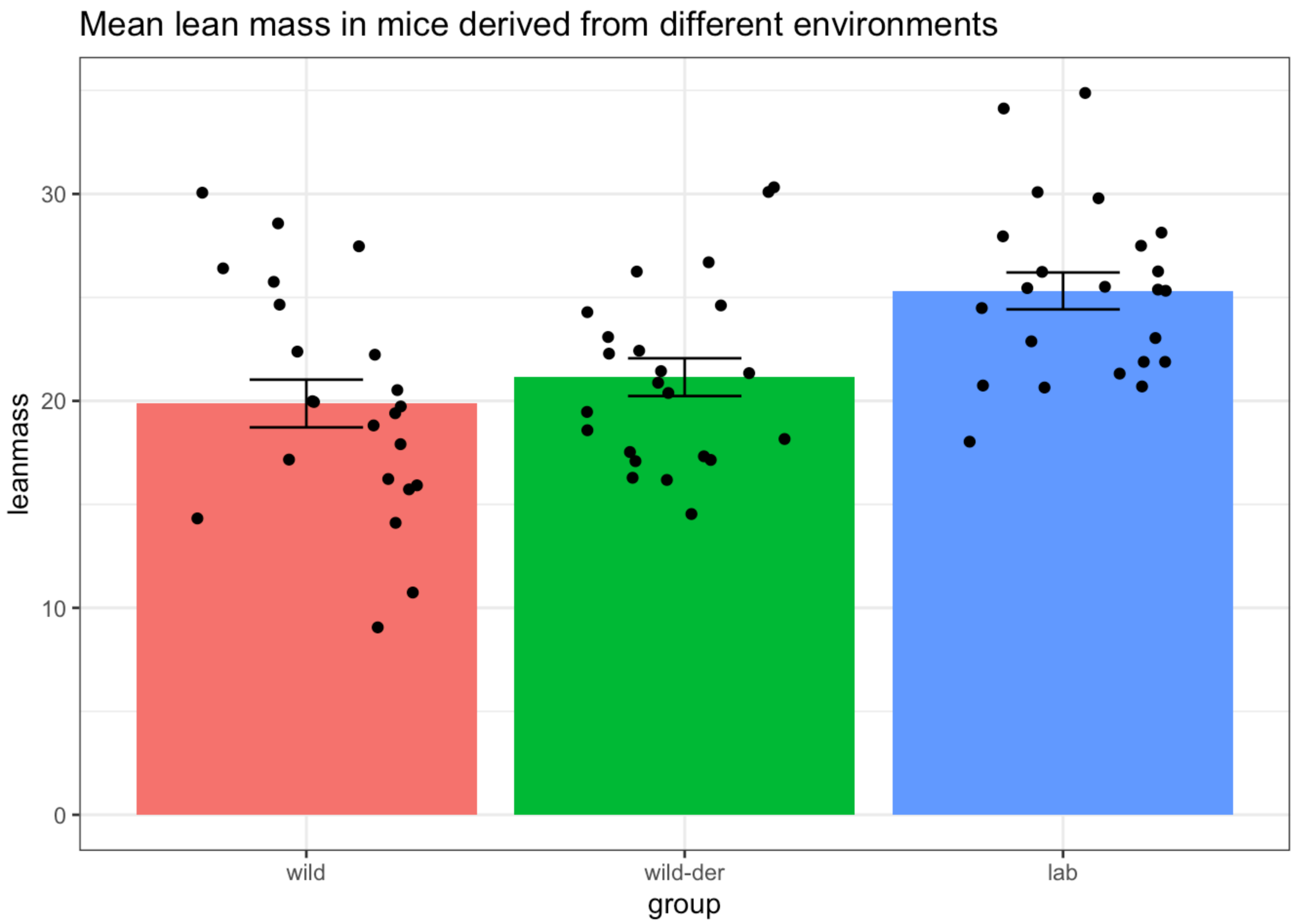
```
summary(mouse.mass)
```

```
##      idnum         group       leanmass         bodyfat
##  1      : 1   wild    :23   Min.   : 9.056   Min.   : 0.1383
##  2      : 1   wild-der:23   1st Qu.:18.160   1st Qu.: 6.0357
##  3      : 1   lab     :23   Median :21.887   Median : 9.0659
##  4      : 1                 Mean   :22.112   Mean   : 9.3414
##  5      : 1                 3rd Qu.:25.755   3rd Qu.:13.1882
##  6      : 1                 Max.   :34.880   Max.   :19.7715
##  (Other):63
```

```
ggplot(data = mouse.mass, mapping = aes(x = group, y = bodyfat)) +
  geom_bar(mapping = aes(fill = group), stat = "summary", fun = "mean", show.legend = FA
LSE) +
  geom_errorbar(stat = "summary", fun.data = "mean_se", width = 0.3) +
  geom_jitter(width = 0.3) +
  labs(title = "Mean body fat in mice derived from different environments")
```



```
ggplot(data = mouse.mass, mapping = aes(x = group, y = leanmass)) +
  geom_bar(mapping = aes(fill = group), stat = "summary", fun = "mean", show.legend = FA
LSE) +
  geom_errorbar(stat = "summary", fun.data = "mean_se", width = 0.3) +
  geom_jitter(width = 0.3) +
  labs(title = "Mean lean mass in mice derived from different environments")
```



Assumptions of ANCOVA

ANCOVA has the same assumptions as ANOVA, along with two extras:

1. Independence
2. Normality of the model residuals within each group
3. Homogeneity of variance of the model residuals across groups
4. Homogeneity of regression slopes (ANCOVA only)
5. Linearity (ANCOVA only)

Independence

We are only measuring each variable once from each mouse and each mouse is only a part of a single group, therefore the independence assumption is met.

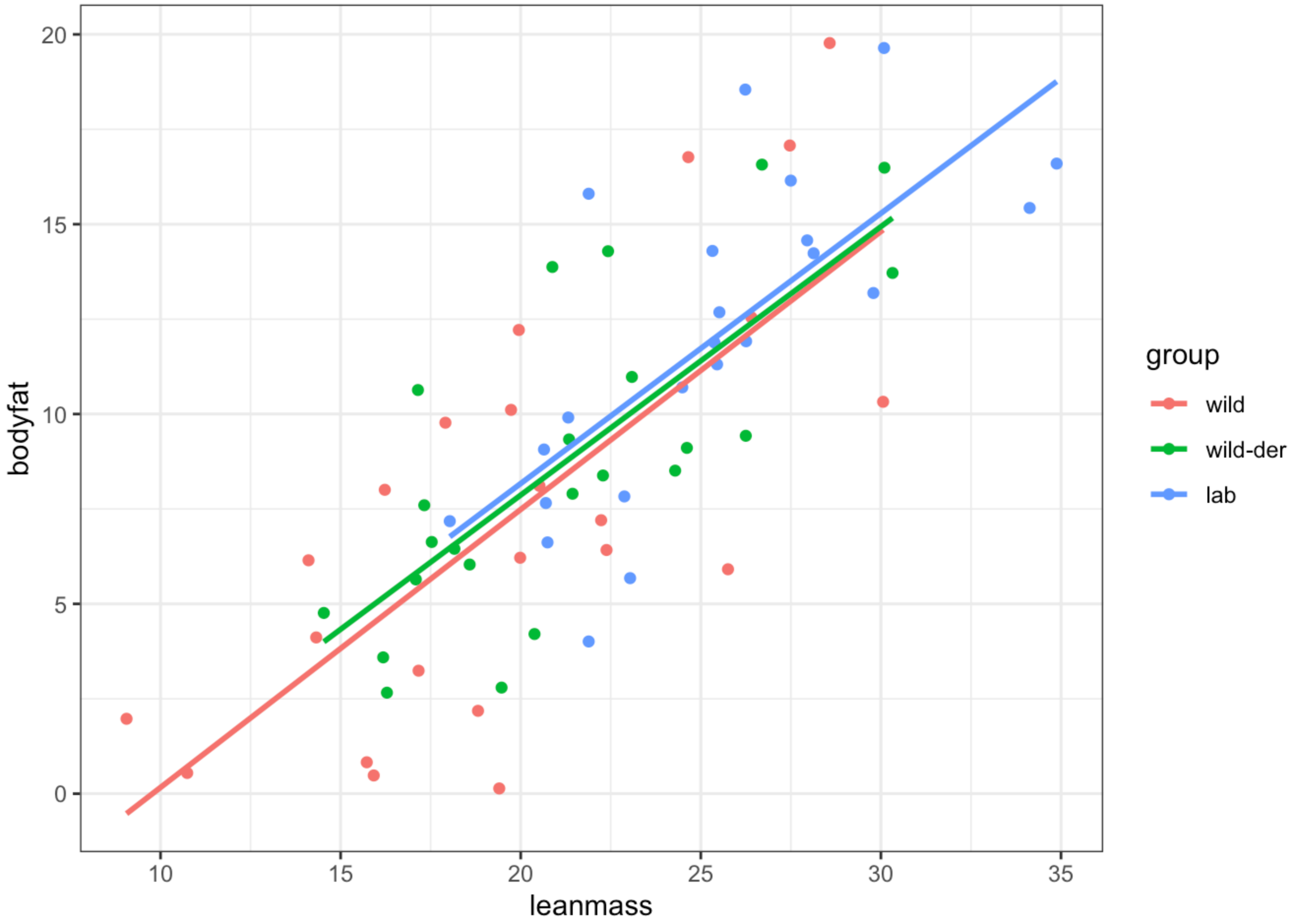
Normality and Homogeneity of Variance

These assumptions will be tested after running the model.

Homogeneity of Regression Slopes and Linearity

The homogeneity of regression slopes and linearity assumptions can be assessed simultaneously. Technically, simple ANOVA is a linear model yet we don't test for linearity. When doing an ANCOVA, part of this linear model tests the relationship between the covariate and the dependent variable within each group. You don't want the relationship between the covariate and the outcome variable to differ between groups. This means that the regression slope within each group should be linear and each slope should be essentially the same. We can test this graphically by plotting scatterplots of the covariate and dependent variable along with regression lines, like we do when testing linearity, except we split the plot by group using the `color` aesthetic.

```
ggplot(data = mouse.mass, aes(x = leanmass, y = bodyfat, color = group)) +
  geom_point() +
  geom_smooth(method = 'lm', se = FALSE)
```



What you want to see is that the slopes for the different groups are essentially parallel, and the data should exhibit a linear relationship. In this case, all three regression slopes are very close to parallel and the relationship seems to be linear, so both assumptions appear to be met.

We can further test the homogeneity of regression slopes assumption mathematically with standardized regression coefficients. We can use the `lm.beta` function from the `QuantPsyc` package to see what the standardized slope is for each group. In general, you do not want any single standardized slope to be 0.4 units away from any other slope.

For this example, we will need to make 3 linear models. Each of them will be testing bodyfat as a function of leanmass within a group. We can filter the dataset and then pipe it into `lm` as follows:

```
mouse.mass %>%
  filter(group == 'wild') %>%
  lm(bodyfat ~ leanmass, data = .) %>%
  lm.beta()
```

```
##  leanmass
## 0.7240212
```

```
mouse.mass %>%
  filter(group == 'wild-der') %>%
  lm(bodyfat ~ leanmass, data = .) %>%
  lm.beta()
```

```
## leanmass
## 0.747509
```

```
mouse.mass %>%
  filter(group == 'lab') %>%
  lm(bodyfat ~ leanmass, data = .) %>%
  lm.beta()
```

```
##  leanmass
## 0.7205002
```

As you can see, all three standardized slopes are between 0.72 and 0.74, which means that none are even close to 0.4 apart. Therefore we have confirmed that the homogeneity of regression slopes assumption is met.

Running an ANCOVA

In order to help your interpretation, we will first run a simple ANOVA, then we will run the ANCOVA by adding the covariate and observe the difference in the models. We will run the regular ANOVA with `ezANOVA` as we did in the last lab. The ANCOVA requires us to use a slightly different version of `ezANOVA` from the `rstatix` package called `anova_test`. It has the exact same inputs, but allows us to additionally adjust for a covariate. The residuals need to be extracted in a slightly different manner however, and it doesn't automatically test homogeneity of variance.

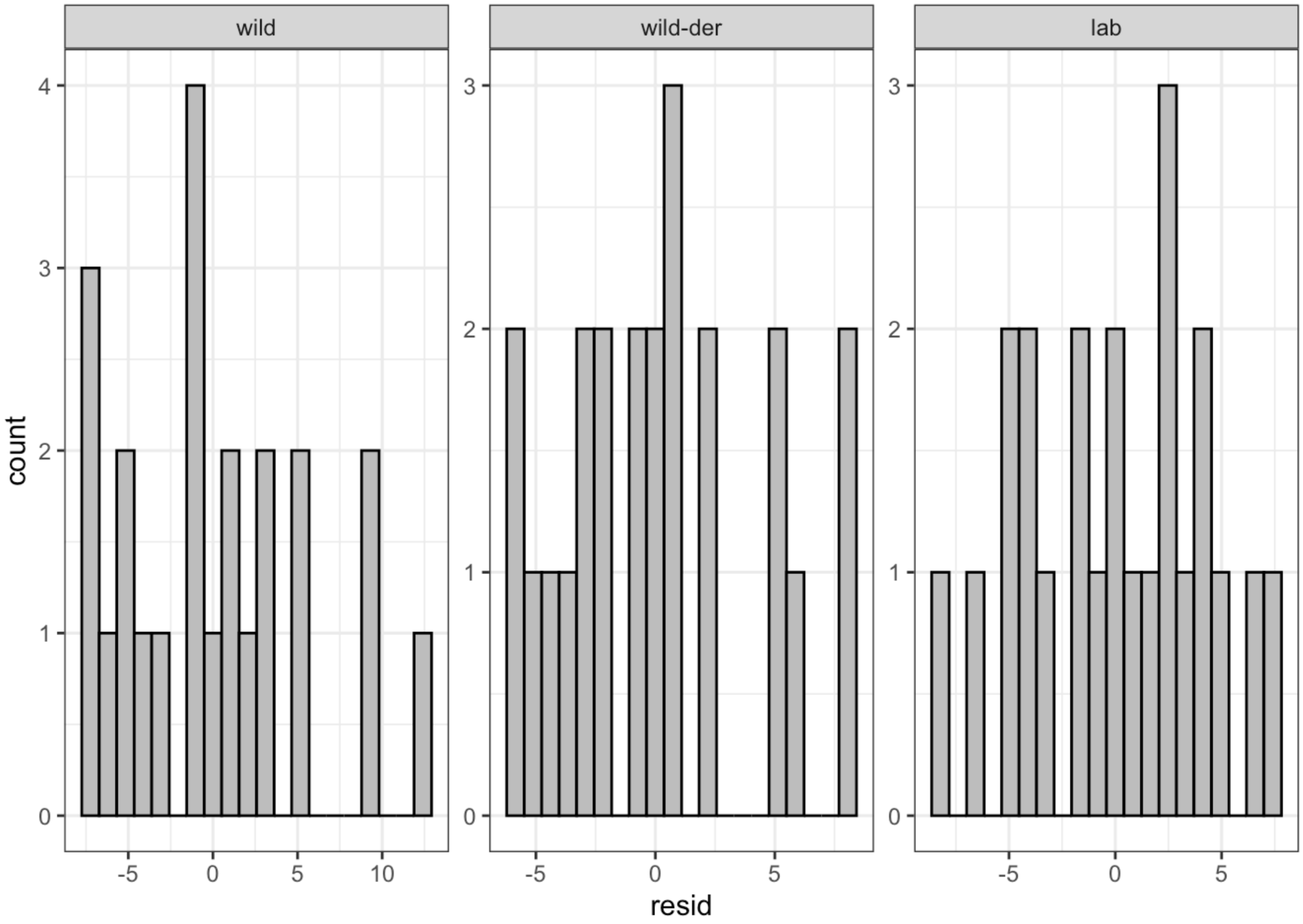
Regular ANOVA

We will both run the basic ANOVA and quickly test it's assumptions below:

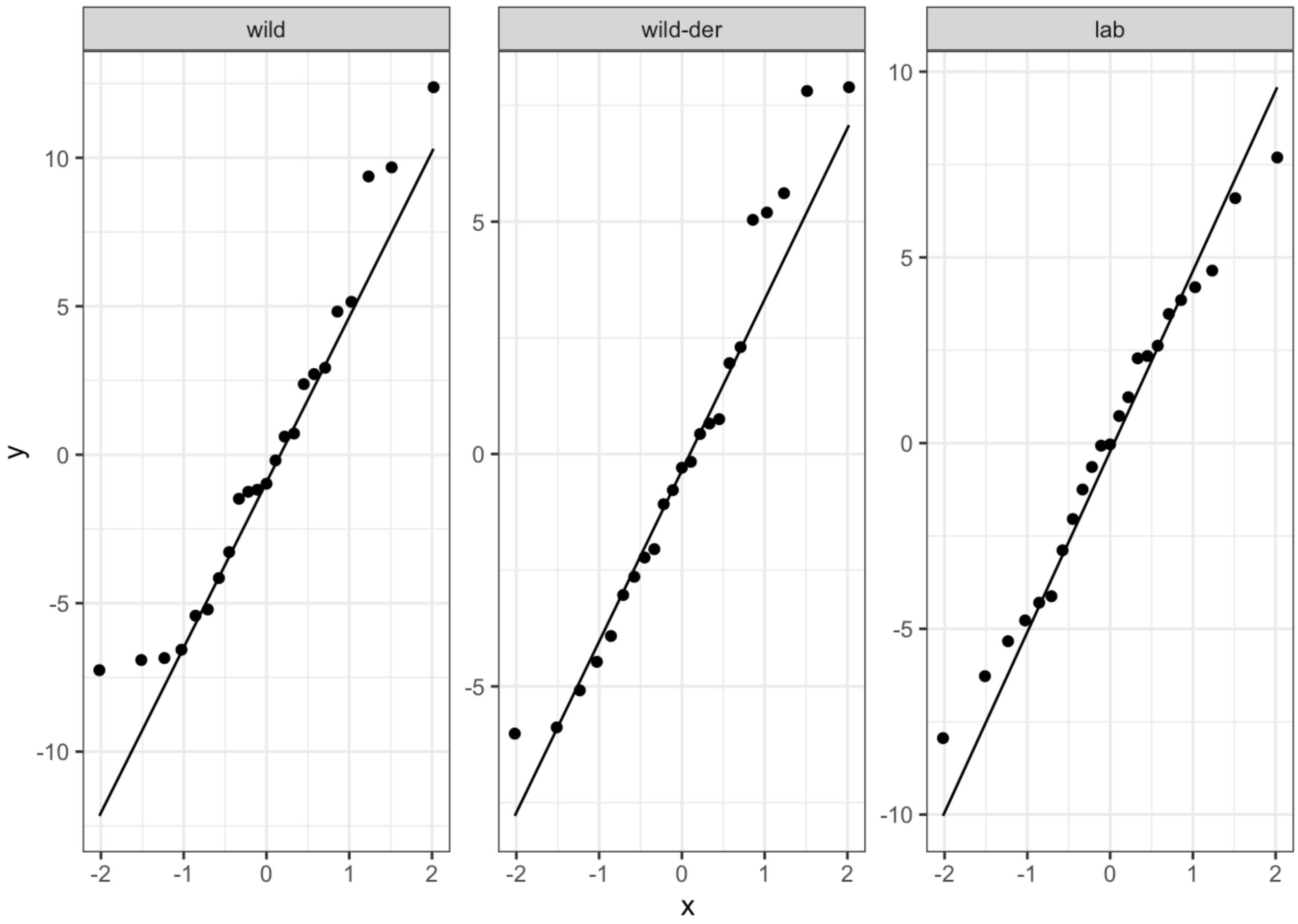
```
mod <- ezANOVA(data = mouse.mass,
               dv = bodyfat,
               between = group,
               wid = idnum,
               type = 3,
               return_aov = TRUE)

mod
```

```
## $ANOVA
##   Effect DFn DFd        F           p p<.05       ges
## 2  group   2  66 5.754132 0.004971247     * 0.1484779
##
## $`Levene's Test for Homogeneity of Variance`
##   DFn DFd      SSn      SSd         F        p p<.05
## 1   2  66 15.98221 529.2673 0.9964964 0.374656
##
## $aov
## Call:
##    aov(formula = formula(aov_formula), data = data)
##
## Terms:
##                     group Residuals
## Sum of Squares   254.0251 1456.8363
## Deg. of Freedom         2        66
##
## Residual standard error: 4.698221
## Estimated effects may be unbalanced
```

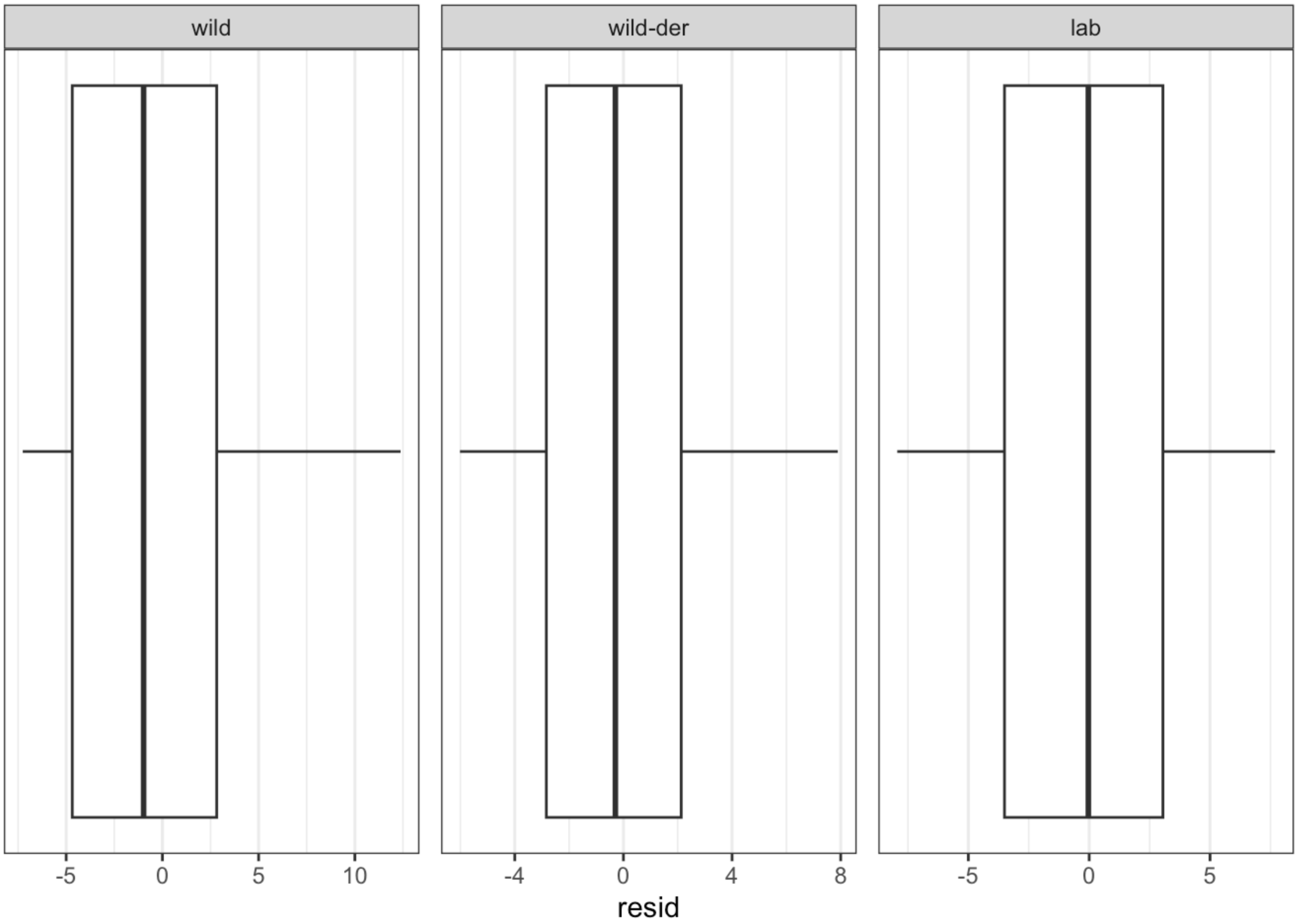
```
# Extracting residuals
residuals <- tibble(group = mouse.mass$group,
                    resid = resid(mod$aov))

# Normality within groups
ggplot(data = residuals, mapping = aes(x = resid)) +
  geom_histogram(bins = 20, 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 wild       0.547    0.568    -0.667   -0.357      0.939      0.171
## 2 wild-der   0.407    0.423    -0.924   -0.494      0.949      0.276
## 3 lab       -0.0768  -0.0798   -1.05    -0.563      0.981      0.915
```

First, we can confirm that our assumptions were met, so this model is valid. We see a significant main effect of group on body fat. This tells us that mean body fat is significantly different between groups. However, there could be a confounding variable that is responsible for this relationship. We can test whether `leanmass` confounds this relationship by adding it as a covariate in an ANCOVA model. Note that in the context of ANCOVA, covariates are often referred to as nuisance variables, as you are not necessarily interested in the effect of the covariate. Instead you are interested in the effect of your predictor variable after adjustment for the covariate.

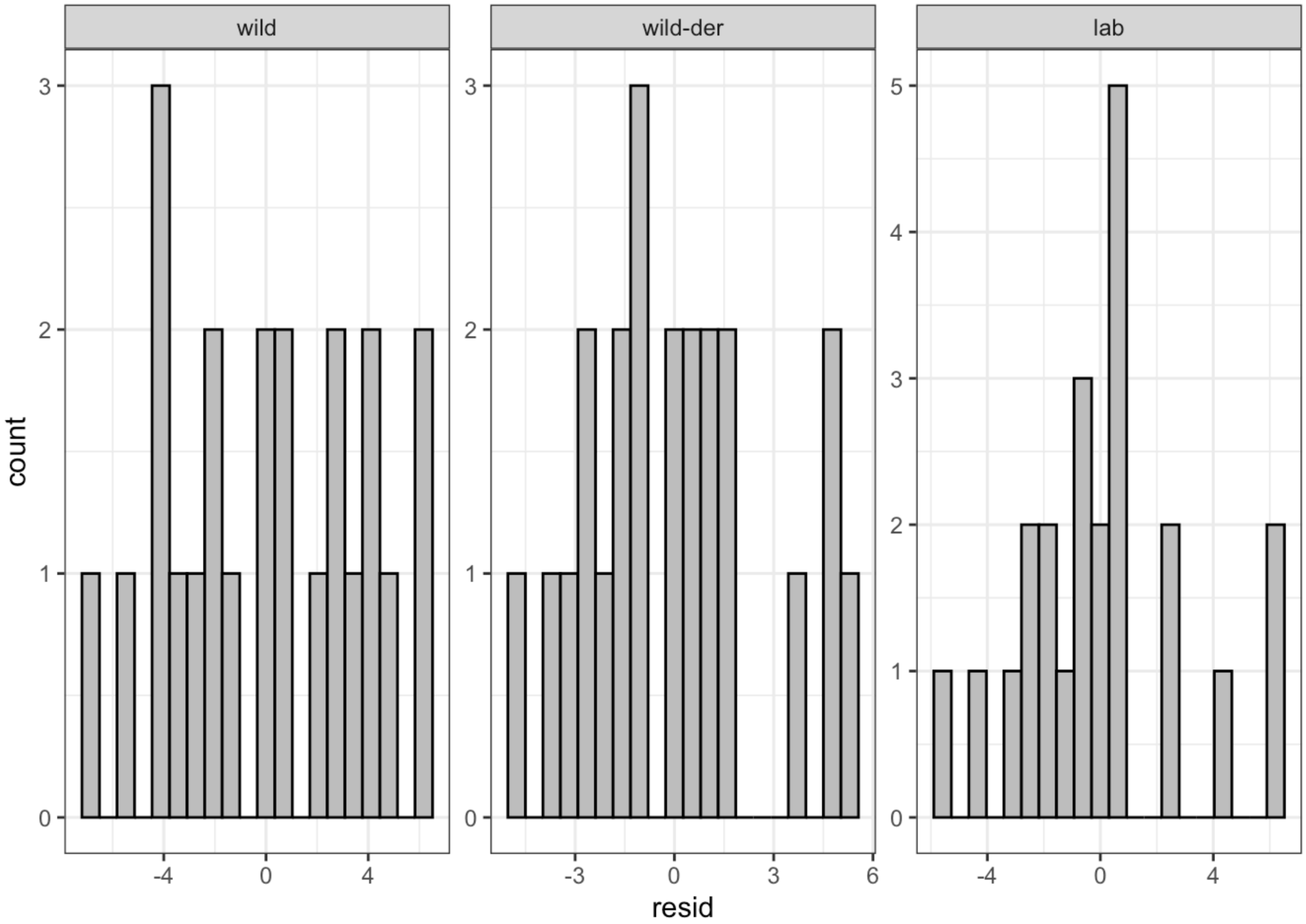
ANCOVA

Let's run an ANCOVA using the `anova_test` method, adding in the `covariate` argument. We can also assess the assumptions as above. Note that rather than using the `mod$aov` to extract residuals as we did above, we will input `attributes(mod)$args$model` . This is just how the anova_test outputs the residuals and unfortunately it won’t output an `aov` object like `ezANOVA` .

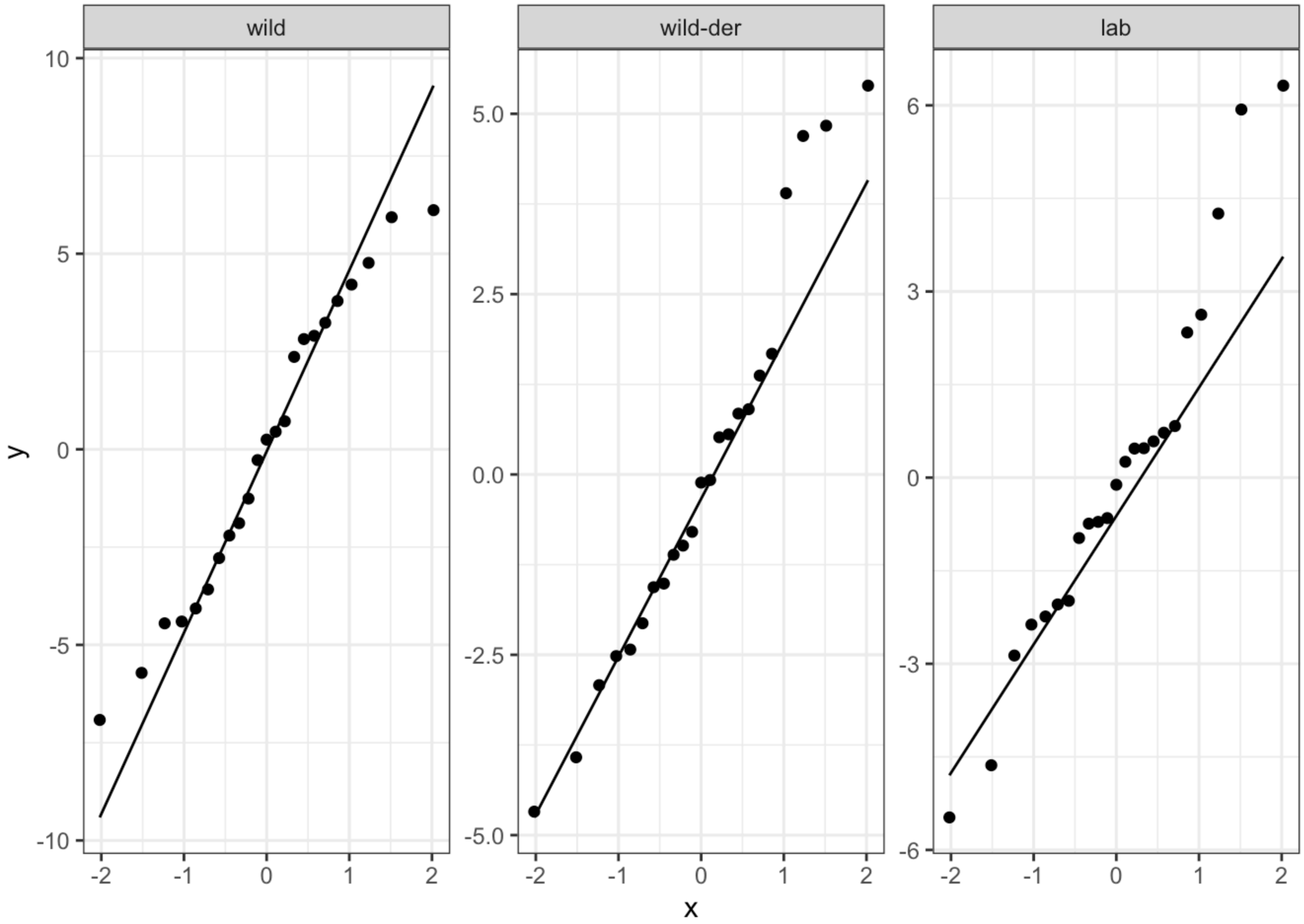
```
mod <- anova_test(data = mouse.mass,
                  dv = bodyfat,
                  between = group,
                  covariate = leanmass,
                  wid = idnum,
                  type = 3)

mod
```

```
## ANOVA Table (type III tests)
##
##     Effect DFn DFd      F        p p<.05   ges
## 1 leanmass   1  65 73.782 2.63e-12     * 0.532
## 2    group   2  65  0.188 8.29e-01       0.006
```

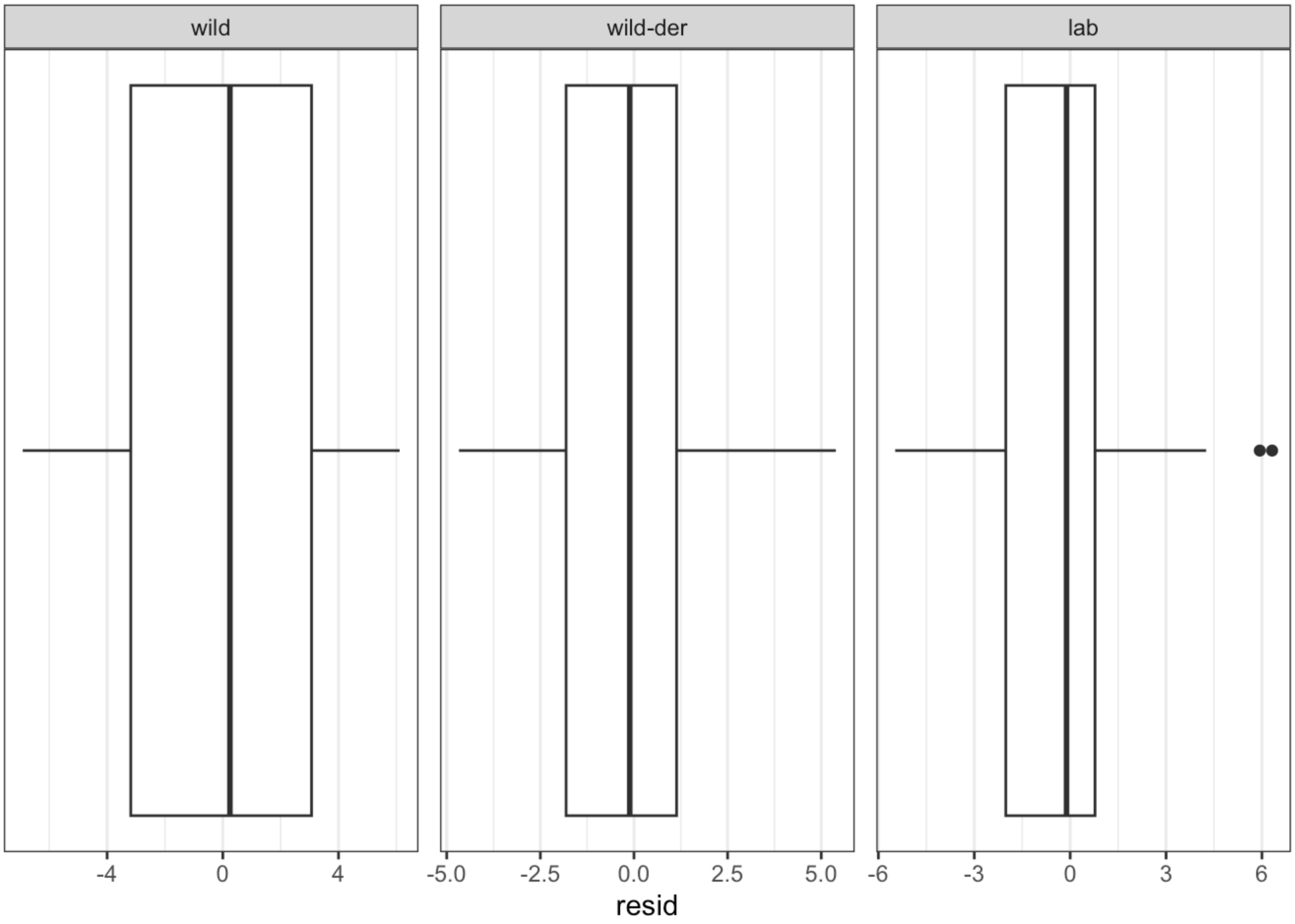
```
# Extracting residuals
residuals <- tibble(group = mouse.mass$group,
                    resid = resid(attributes(mod)$args$model))

# Normality within groups
ggplot(data = residuals, mapping = aes(x = resid)) +
  geom_histogram(bins = 20, 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 wild      -0.0578  -0.0600   -1.31   -0.701       0.961      0.474
## 2 wild-der   0.428    0.445    -0.707  -0.378       0.954      0.352
## 3 lab        0.416    0.432    -0.183  -0.0979      0.957      0.414
```

```
# Homogeneity of variance - wasn't tested by anova_test
leveneTest(resid ~ group, data = residuals)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value  Pr(>F)
## group  2  2.7213 0.07318 .
##       66
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

First, we can confirm that our assumptions were met, so this model is also valid. We can now see that the main effect of the `group` variable has become non-significant in this new model. There is, however, a highly significant effect of the covariate (`leanmass`) instead. So the covariate (`leanmass`) in this case was confounding the relationship between `group` and `bodyfat`, making it seem like `group` had an effect on `bodyfat` when in fact it likely just influences `leanmass` which correlates with `bodyfat`.

Reporting an ANCOVA

You can either report the ANOVA first and then the ANCOVA, or you can just report the ANCOVA alone.

“When not adjusting for lean body mass, there was a significant difference in mean body fat between groups of lab-reared mice, wild mice, and wild-derived mice, F(2,66) = 5.754, p = 0.00497.

After adjusting for lean body mass, the effect of group was no longer significant, F(2,65) = 0.188, p = 0.829.”

Independent Practice

For your independent practice, you will need to load the `mtcars` dataset using the command `mtcars <- mtcars`. In this example, assume `mpg` (miles per gallon) is the outcome variable, `wt` (car weight) is the covariate, and `gear` (number of gears) is our predictor variable. Currently, the `gear` variable is numeric, and thus needs to be transformed into a factor prior to running the model. Recall we did this last week using the `mutate` function along with the `factor` function.

Determine whether your unadjusted model testing `mpg` between car `gear` groups (3, 4, 5) is significant. After adjusting for `wt`, determine whether the model is still significant.

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