R Workshop: ANOVA, ANCOVA & Factorial ANOVA

Brier Gallihugh, M.S.

2023-06-19

Table of contents

Introduction to ANOVA

Running ANOVA

```
library(car)
library(psych)
library(multcomp)
library(effects)
library(tidyverse)
library(sjstats)


data <- dplyr::starwars

data <- data %>%
  dplyr::select(height,mass,hair_color,species,sex) %>%
  na.omit() %>%
  filter(species == "Gungan" | species == "Human" | species == "Wookiee")①

height_species_aov <- aov(height ~ species, data = data)                ②
summary(height_species_aov)                                             ③

# Post Hoc Test (Tukey)
TukeyHSD(height_species_aov)                                            ④
# Bonferroni Adjustment vs Tukey
pairwise.t.test(data$height, data$species, p.adjust.method = "bonferroni")⑤
```

① Here we are using the `filter()` function to filter for "Gungan", "Human" and "Wookiee"

② This is the `aov()` function. Here we are specifying the ANOVA formula.

③ The `summary()` function will give you the output of the ANOVA

④ Because the `species` variable has more than 2 conditions, we might want to figure out which comparisons are statistically different from each other. The `TukeyHSD()` function allows us to perform a post hoc Tukey test

⑤ Maybe Tukey isn't your favorite correction. Another route might be to use a pairwise t test that corrects for multiple comparisons using a Bonferroni correction. You can do that with the `pairwise.t.test()` function.

```
            Df Sum Sq Mean Sq F value   Pr(>F)
species      2   6118    3059    22.5 3.86e-06 ***
Residuals   23   3127     136
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = height ~ species, data = data)

$species
                    diff        lwr       upr     p adj
Human-Gungan   -30.45455 -52.022229 -8.886862 0.0048319
Wookiee-Gungan  21.00000  -8.202782 50.202782 0.1915885
Wookiee-Human   51.45455  29.886862 73.022229 0.0000125


    Pairwise comparisons using t tests with pooled SD

data:  data$height and data$species

        Gungan Human
Human   0.0053 -
Wookiee 0.2545 1.3e-05

P value adjustment method: bonferroni
```

Assumptions of ANOVA

As ANOVA is just a special case of linear regression, the same assumptions exist for ANOVA that exist for linear regression. We will test each assumption below. The code will look awfully similar to the code used for linear regression previously

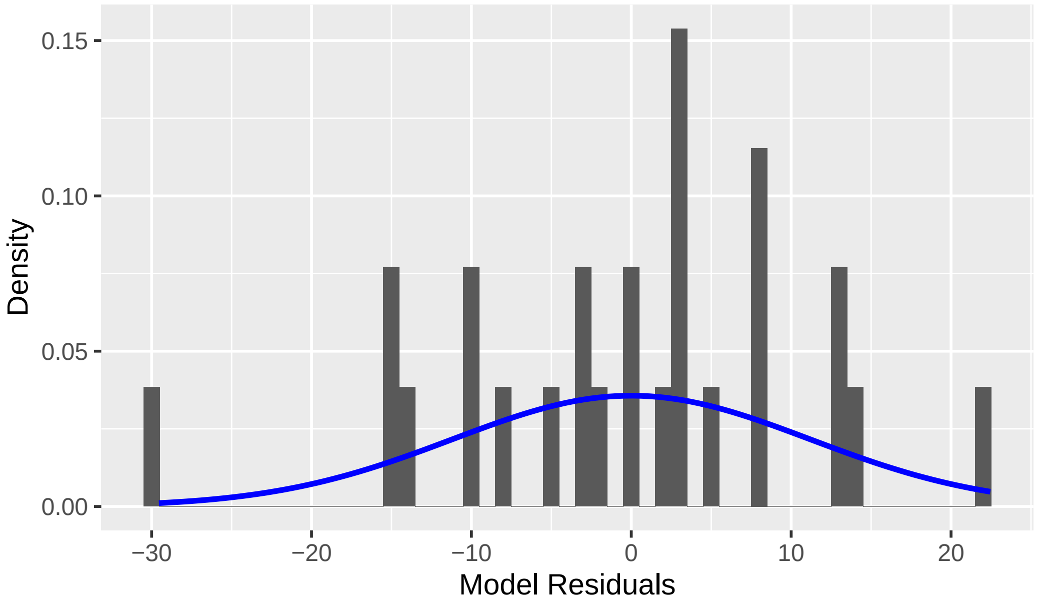
Model Normality

Graphical

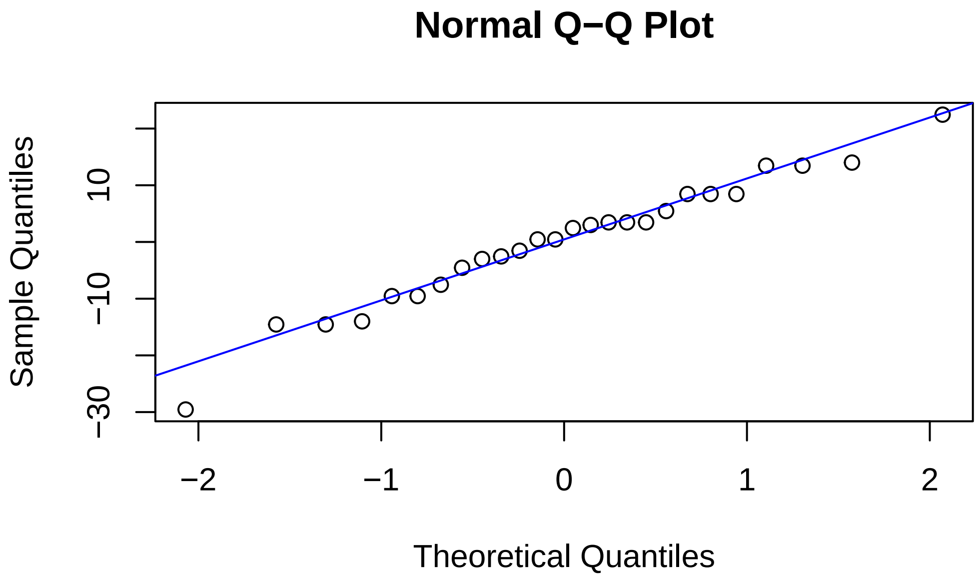
```
density_plot <- data %>% ggplot(aes(x = height_species_aov$residuals)) +
  geom_histogram(aes(y= after_stat(density)),binwidth = 1) +
  stat_function(fun = dnorm,
                args = list(mean = mean(height_species_aov$residuals),
                            sd = sd(height_species_aov$residuals)),
                col = "blue",
                linewidth = 1) +
  labs(title = "Figure 1. Histogram of Model Residual Scores",
```

```
       x = "Model Residuals",
       y = "Density")

density_plot
```

Figure 1. Histogram of Model Residual Scores



```
# OR

qqnorm(height_species_aov$residuals)                               ①
qqline(height_species_aov$residuals, col = "blue")
```



1.When running the `qqnorm()` and `qqline()` functions at the same time, you will get a qq-plot which will graphically assess the normality of the argument given (in this case the model residuals). The `col` argument simply tells R which color to make the reference line.

Statistical

```
psych::describe(height_species_aov$residuals)                          ①
shapiro.test(height_species_aov$residuals)
```

① Another way to assess normality is to look at the skew and kurtosis of the data. Typically skew and kurtosis between -1 and 1 are acceptable. We can see this using the `describe()` function within the `psych` package.

```
   vars  n mean    sd median trimmed   mad    min   max range  skew kurtosis
X1    1 26    0 11.18   1.45    0.35 10.38 -29.55 22.45    52 -0.44     0.16
     se
X1 2.19

    Shapiro-Wilk normality test

data:  height_species_aov$residuals
W = 0.97347, p-value = 0.7143
```

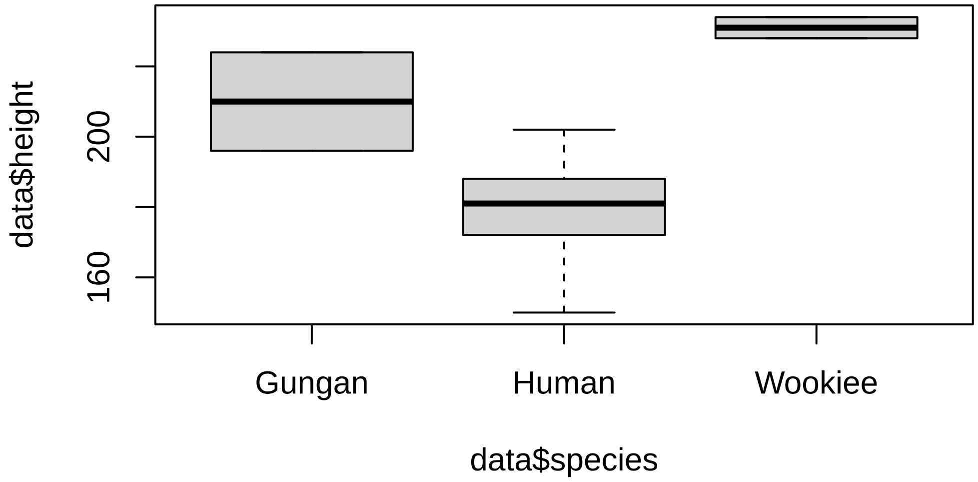
Homogeneity of Variance

As with regression, homogeneity of variance is also important for ANOVA. We will graphically and statistically assess this assumption below.

Graphical

```
boxplot(data$height ~ data$species)                                    ①
```

① The `boxplot()` function will give us a visual representation of the DV at each level of the IV. We are looking for boxplots that are roughly the same height across each group.



Statistical

Statistically, we can assess homogeneity of variance using Levene's Test. Here we want the results to not be statistically significant.

```
leveneTest(data$height ~ data$species)                                 ①
```

① The `leveneTest()` function from the `car` package takes a DV and IV as arguments.

```
Levene's Test for Homogeneity of Variance (center = median)
      Df F value Pr(>F)
group  2   1.168 0.3288
      23
```

Introduction to ANCOVA

Running ANCOVA

ANCOVA refers to any ANOVA model with 2 or more parameters. These models require similar assumptions to ANOVA with a couple of additional ones. Below we will see how to run an ANCOVA and then test its assumptions.

```
data <- data %>% mutate(species = as.factor(species))
# Type I SS
ancova <- aov(height ~ species + mass, data = data)
summary(ancova)                                                         ①
# Reversed Order
ancova2 <- aov(height ~ mass + species, data = data)
summary(ancova2)                                                        ②

car::Anova(ancova2, type = "III")                                       ③

# Std Means
summary_data <- data %>%
  dplyr::group_by(species) %>%
  dplyr::summarise(n = n(),
                   mean = mean(height))

# Adjust means for covariate effect
adjustedMeans<- effect("species", ancova, se=TRUE)                      ④
summary(adjustedMeans)
adjustedMeans$se

# Post Hoc Tests

posthoc <- multcomp::glht(ancova, linfct = multcomp::mcp(species = "Tukey"))⑤
summary(posthoc)
confint(posthoc)

# Effect size
```

```
sjstats::anova_stats(car::Anova(ancova,type = "III"))                    ⑥
```

① This displays our ANCOVA output with species and then mass.
② This displays our ANCOVA output with mass and then species. We can see that 1 and 2 are different. This is because R by default does a sequential ANCOVA so order matters.
③ To get around this, we can use the `Anova()` function in the `car` package to specify that we want Type III sums of squares. This will then generate an ANCOVA where the order of predictors doesn't matter
④ When having covariates, one might wish to report adjusted means. We can do that using the `effect()` function from the `effects` package.
⑤ To run a post hoc analysis, we need to use the `ghlt()` function from the `multcomp` package. There are additional corrections you can do outside of Tukey. To investigate, please see the documentation.
⑥ To get basic ANOVA effect size statistics, we can use the `anova_stats()` function from the `sjstats` package.

```
            Df Sum Sq Mean Sq F value   Pr(>F)    
species      2   6118  3059.0   37.27 8.61e-08 ***
mass         1   1322  1321.8   16.10 0.000584 ***
Residuals   22   1806    82.1                     
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
            Df Sum Sq Mean Sq F value   Pr(>F)    
mass         1   3737    3737   45.53 8.81e-07 ***
species      2   3702    1851   22.55 4.70e-06 ***
Residuals   22   1806      82                     
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Anova Table (Type III tests)

Response: height
             Sum Sq Df F value    Pr(>F)    
(Intercept) 28097.6  1 342.337 6.734e-15 ***
mass         1321.8  1  16.104 0.0005843 ***
species      3702.4  2  22.555 4.699e-06 ***
Residuals    1805.7 22                      
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

 species effect
species
  Gungan    Human  Wookiee 
```

```
214.4986 180.5408 215.5526

 Lower 95 Percent Confidence Limits
species
  Gungan    Human  Wookiee
201.0113 176.5022 200.0533

 Upper 95 Percent Confidence Limits
species
  Gungan    Human  Wookiee
227.9859 184.5794 231.0520
[1] 6.503427 1.947367 7.473624

     Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: Tukey Contrasts


Fit: aov(formula = height ~ species + mass, data = data)

Linear Hypotheses:
                      Estimate Std. Error t value Pr(>|t|)
Human - Gungan == 0    -33.958      6.748  -5.033 0.000166 ***
Wookiee - Gungan == 0    1.054     10.333   0.102 0.993965
Wookiee - Human == 0    35.012      7.846   4.462 0.000476 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Adjusted p values reported -- single-step method)


     Simultaneous Confidence Intervals

Multiple Comparisons of Means: Tukey Contrasts


Fit: aov(formula = height ~ species + mass, data = data)

Quantile = 2.4798
95% family-wise confidence level


Linear Hypotheses:
                      Estimate lwr      upr
```

```
Human - Gungan == 0   -33.9578 -50.6904 -17.2252
Wookiee - Gungan == 0   1.0541 -24.5705  26.6786
Wookiee - Human == 0   35.0118  15.5560  54.4677
```

```
term      |    sumsq |   meansq | df | statistic | p.value | etasq | partial.etasq | omegasq
-------------------------------------------------------------------------------------------
species   | 3702.445 | 1851.223 |  2 |    22.555 |  < .001 | 0.542 |         0.672 |   0.512
mass      | 1321.785 | 1321.785 |  1 |    16.104 |   0.001 | 0.194 |         0.423 |   0.179
Residuals | 1805.669 |   82.076 | 22 |           |         |       |               |
```

Assumptions of ANCOVA

ANCOVA requires the same assumptions as ANOVA with two additional assumptions: 1) That the predictor and covariate are independent 2) The regression slopes are homogeneous

We will test each of the traditional parametric tests as well as the two additional assumptions below.

Predictor x Covariate Indepenence

Statistical

```r
predictor_assumption <- aov(mass ~ species, data = data)          ①
summary(predictor_assumption)
```

① To assess this statistically, one needs to run an ANOVA looking at each predictor with each additional parameter. We do not want these to be statistically significant.

```
            Df Sum Sq Mean Sq F value Pr(>F)
species      2   3390  1695.1   4.694 0.0195 *
Residuals   23   8306   361.1
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

> **Tip**
>
> Here the statistical test for this assumption is statistically significant (which is bad). It actually means we can’t do this (unless we have random assignment)

Homogeneity of Regression Slopes

Statistical

```
regression_slope_assumption <- aov(height ~ species*mass, data = data)
car::Anova(regression_slope_assumption, type = "III")                    ①
```

① To test this assumption, one just needs to test the interaction effects within the model. The code for this is shown here.

```
Anova Table (Type III tests)

Response: height
              Sum Sq Df F value  Pr(>F)
(Intercept)   149.72  1  1.9153 0.18163
species       155.01  2  0.9914 0.38856
mass          392.00  1  5.0144 0.03666 *
species:mass  242.18  2  1.5490 0.23689
Residuals    1563.48 20
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Tip

For the above example, the interaction is NOT significant so assumption is met

Residual Normality (DV ~ IV)

Residual normality is an assumption shared with standard regression. We can assess it the same way we did previously when looking at simple ANOVA. The same is true for the homogeneity of variance assumption.

Statistical

```
height_mass_resid <- scale(residuals(aov(height ~ mass, data = data))) ①
psych::describe(height_mass_resid)                                     ②
shapiro.test(height_mass_resid)                                        ③
```

① The `scale()` function will scale the residuals of the ANOVA. The `residuals()` function pulls out the residuals within the ANOVA

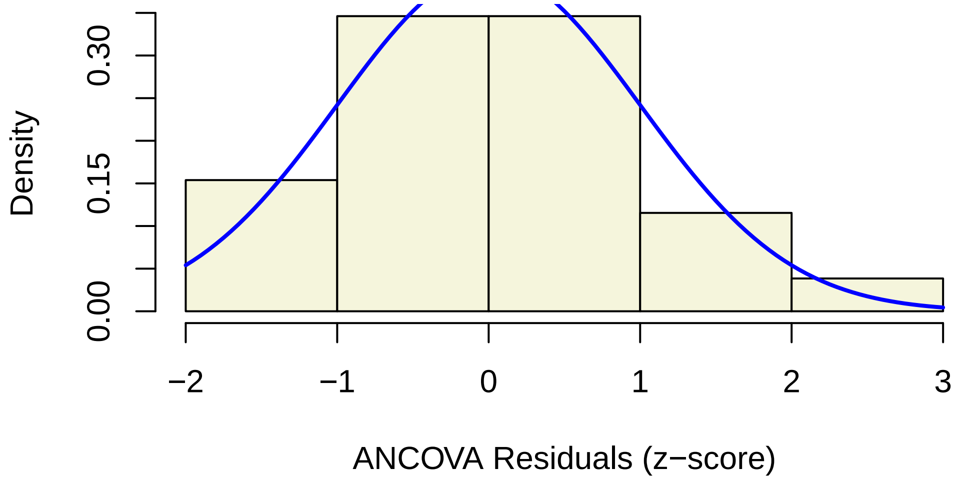
② Statistical summary of the residuals using the `describe()` function in the `psych` package
③ Statistical test (Shapiro Wilks) of the residuals using the `shapiro.test()` function

```
   vars  n mean sd median trimmed  mad   min max range skew kurtosis  se
X1    1 26    0  1      0   -0.06 0.82 -1.85 2.7  4.55 0.71     0.44 0.2

    Shapiro-Wilk normality test

data:  height_mass_resid
W = 0.95143, p-value = 0.2505
```

Graphical

```
hist(height_mass_resid, col = 'beige',                                     ①
     main="", xlab = "ANCOVA Residuals (z-score)",
     probability = TRUE)
curve(dnorm(x, mean = mean(height_mass_resid),
            sd = sd(height_mass_resid)),
            add = TRUE, lwd = 2, col = 'blue')
```

① A graphical histogram (with normal distribution overlay) of the model residuals



Homogeneity of Variance

Statistical

```
# Levene's Test to Assess Equal Variance for Species
car::leveneTest(data$height ~ data$species)                      ①
```

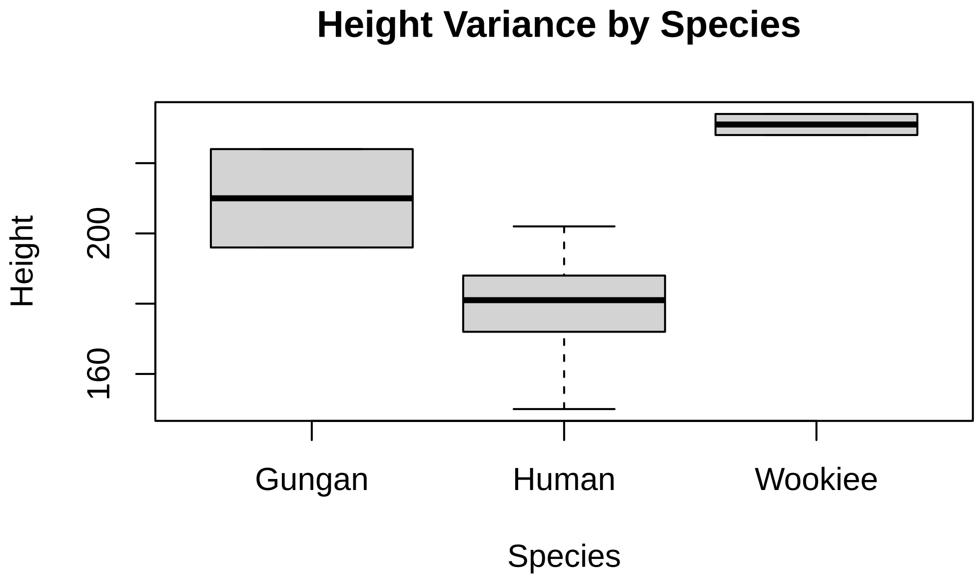
① A statistical test (Levene's test) of the homogeneity assumption using the `leveneTest()` function in the `car` package.

```
Levene's Test for Homogeneity of Variance (center = median)
      Df F value Pr(>F)
group  2   1.168 0.3288
      23               
```

Graphical

```
# Boxplot Height by Species
boxplot(height ~ species,data=data, main="Height Variance by Species ",①
   xlab="Species", ylab="Height")
```

① A visual representation of the homogeneity of variance assumption using a boxplot



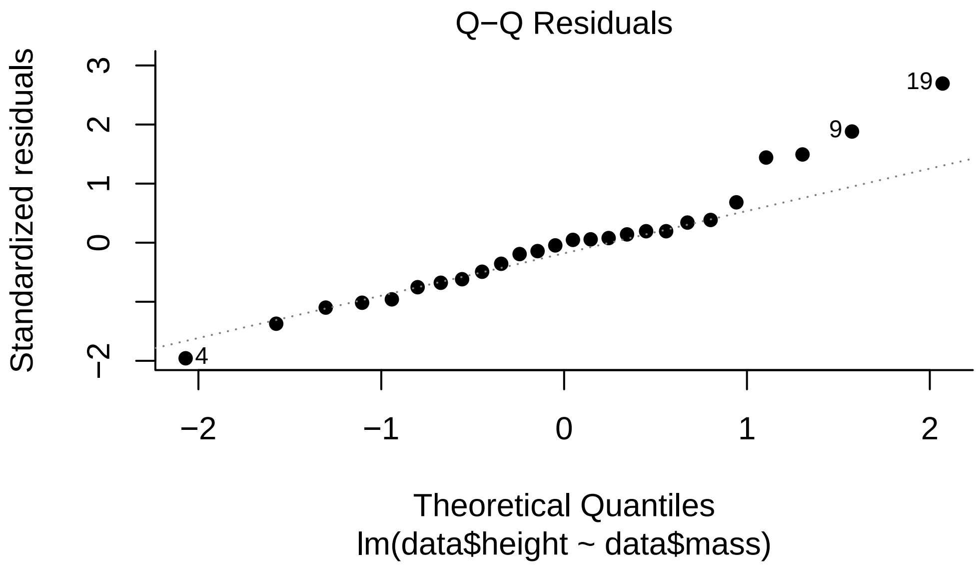
Linearity of CV & DV

The last assumption is that the CV and DV are linearly related. We can test this using the code below graphically.

Graphical

```
plot(lm(data$height ~ data$mass,data = data),                                 ①
     pch = 16, bty = 'l',2)
```

① Visual representation of the CV to DV linearity assumption



Introduction to Factorial ANOVA

Running Factorial ANOVA

```
library(tidyverse)
factorial_data <- starwars %>%                                                ①
  select(mass,homeworld,species) %>%                                          ②
  mutate(homeworld = as.factor(homeworld),                                    ③
```

```
         species = as.factor(species)) %>%                                ④
  filter(homeworld == "Tatooine" | homeworld == "Naboo") %>%                ⑤
  filter(species == "Human" | species == "Droid") %>% na.omit()             ⑥
# Modeling As Factorial ANOVA
aov_factorial <- aov(mass ~ species*homeworld, data = factorial_data)       ⑦
Anova(aov_factorial, type = "III")                                          ⑧

# Modeling As A Regression (For SS Analyses)
reg_fanova <- lm(mass ~ homeworld*species,data = factorial_data)            ⑨
summary(reg_fanova)# <10>
```

① Start with the `starwars` data set
② Use the `select()` function to pull out mass, homeworld and specices variables
③ Call the `mutate()` function to format the homeworld variable using the `as.factor()` function.
④ Format the species variables as a factor using the `as.factor()` function.
⑤ Use the `filter()` function to filter for observations with homeworld = “Tatooine” OR “Naboo”
⑥ Use the `filter()` function to filter for observations with species = “Human” OR “Droid”
⑦ Create an ANOVA object using the `aov()` function
⑧ Use the `Anova()` function from the `car()` package to specifiy Type III Sums of Squares
⑨ Create an Regression object using the `lm()` function
①**0** Show output of the regression object using the `summary()` function.

```
Anova Table (Type III tests)

Response: mass
                  Sum Sq Df F value Pr(>F)
(Intercept)       1024.0  1  1.6199 0.2388
species            990.1  1  1.5662 0.2461
homeworld          308.2  1  0.4875 0.5048
species:homeworld   19.0  1  0.0301 0.8666
Residuals         5057.2  8

Call:
lm(formula = mass ~ homeworld * species, data = factorial_data)

Residuals:
   Min     1Q Median     3Q    Max
-23.33 -19.50  -6.00  17.88  40.00

Coefficients:
```

```
                               Estimate Std. Error t value Pr(>|t|)
(Intercept)                      32.000     25.143   1.273    0.239
homeworldTatooine                21.500     30.793   0.698    0.505
speciesHuman                     36.333     29.032   1.251    0.246
homeworldTatooine:speciesHuman    6.167     35.557   0.173    0.867

Residual standard error: 25.14 on 8 degrees of freedom
Multiple R-squared:  0.5219,    Adjusted R-squared:  0.3426 
F-statistic:  2.91 on 3 and 8 DF,  p-value: 0.1009
```

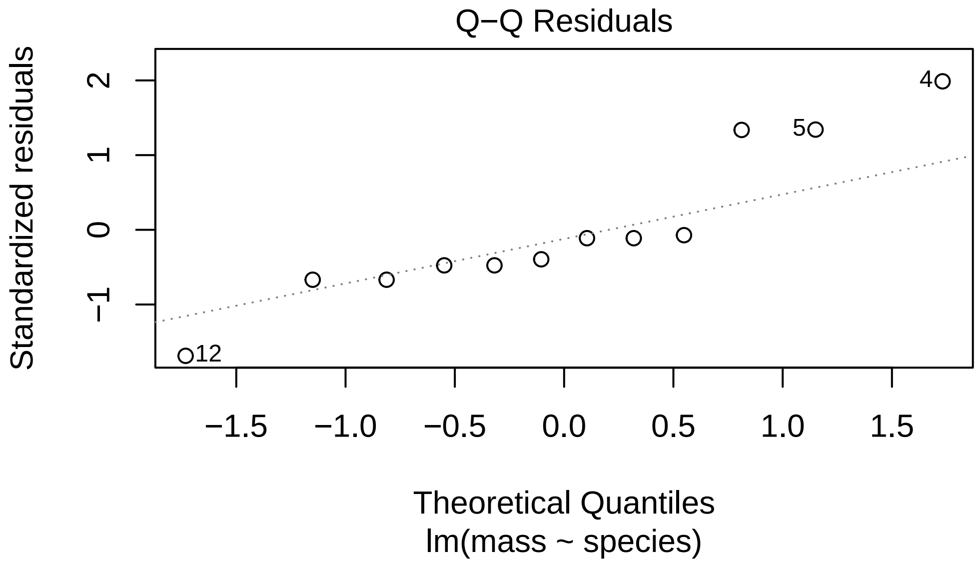
Assumptions of Factorial ANOVA

Linearity of IV to DV (Regression)

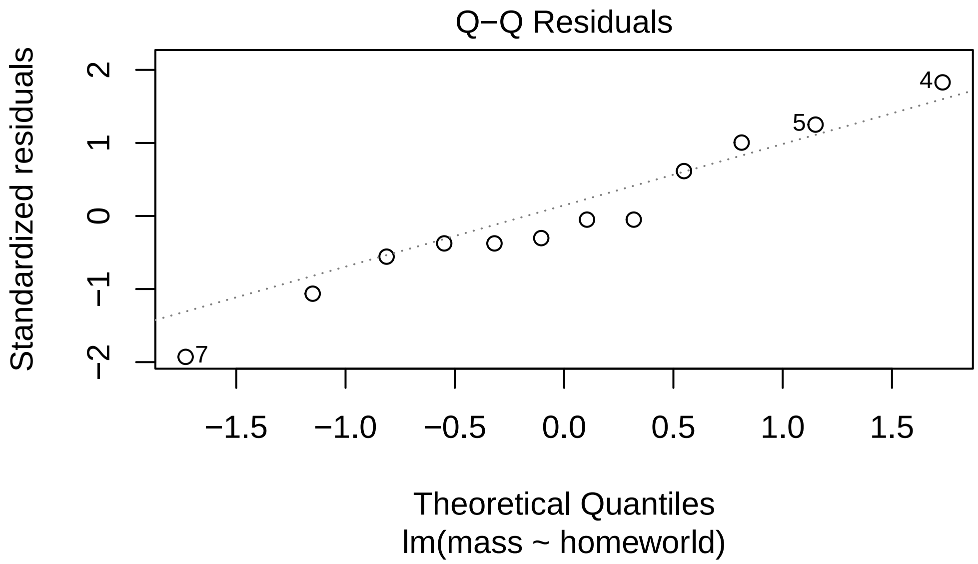
Graphical

```
plot(lm(mass ~ species, data = factorial_data),2)                      ①
```

② Plot the residuals of the model using the `plot()` and `lm()` functions for mass on homeworld



```r
plot(lm(mass ~ homeworld, data = factorial_data),2)                    ②
```



Statistical

```r
psych::describe(resid(aov_factorial))                                  ①
shapiro.test(resid(aov_factorial))                                     ②
```

① Statistical summary of the model residuals using the `describe()` function in the `psych` package

② Statistical test (Shapiro Wilk) of the residuals using the `resid()` and `shapiro.test()` functions

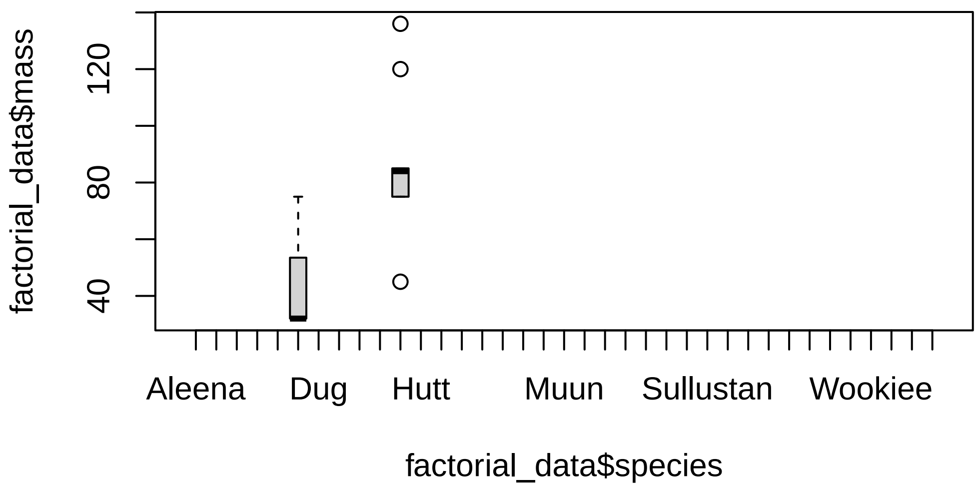
```
   vars  n mean    sd median trimmed   mad    min max range skew kurtosis   se
X1    1 12    0 21.44     -6   -1.67 22.61 -23.33  40 63.33 0.43     -1.4 6.19

    Shapiro-Wilk normality test

data:  resid(aov_factorial)
W = 0.90166, p-value = 0.1666
```

Homogeneity of Variance (Regression)

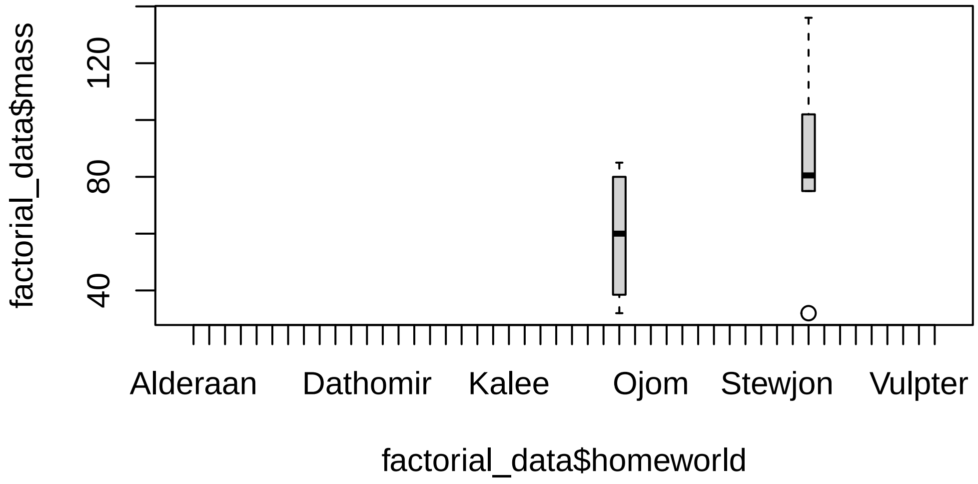
Graphical

```
boxplot(factorial_data$mass ~ factorial_data$species)                 ①
```

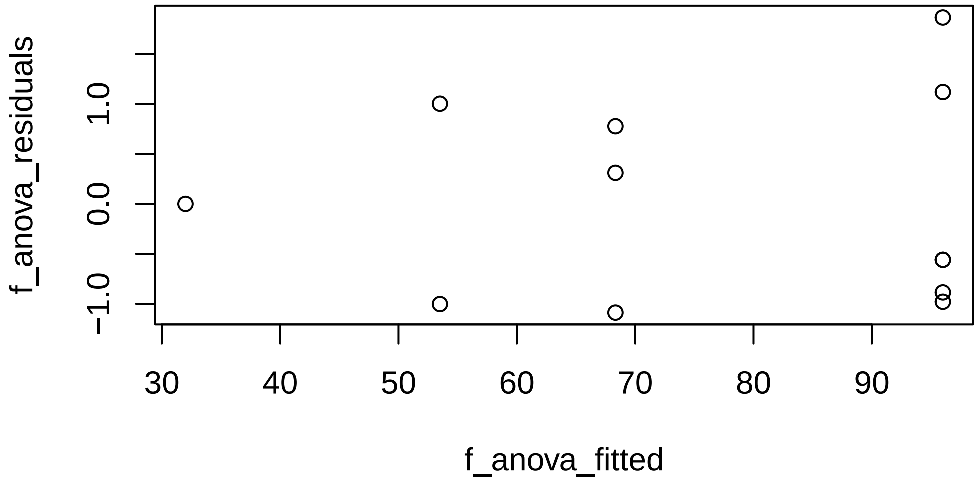
③ Fitted object of the ANCOVA
④ Plot of the residual x fitted data using the `plot()` function



```
boxplot(factorial_data$mass ~ factorial_data$homeworld)               ①
```



```r
# Residuals
f_anova_residuals <- scale(residuals(lm(mass ~ species*homeworld, data = factorial_data)))
f_anova_fitted <- fitted(lm(mass ~ species*homeworld, data = factorial_data))

plot(f_anova_residuals ~ f_anova_fitted)                                    ④
```



Statistical

```
leveneTest(factorial_data$mass ~ factorial_data$species)                          ①

leveneTest(factorial_data$mass ~ factorial_data$homeworld)                        ②

leveneTest(factorial_data$mass ~ interaction(factorial_data$homeworld,factorial_data$speci
```

① Levene test of mass x species using `leveneTest()` function
② Levene test of mass x homeworld using `leveneTest()` function
③ Levene test of mass x species/homeworld interaction using `leveneTest()` function

```
Levene's Test for Homogeneity of Variance (center = median)
      Df F value Pr(>F)
group  1  0.0367  0.852
      10               
Levene's Test for Homogeneity of Variance (center = median)
      Df F value Pr(>F)
group  1   1e-04 0.9919
      10               
Levene's Test for Homogeneity of Variance (center = median)
      Df F value Pr(>F)
```

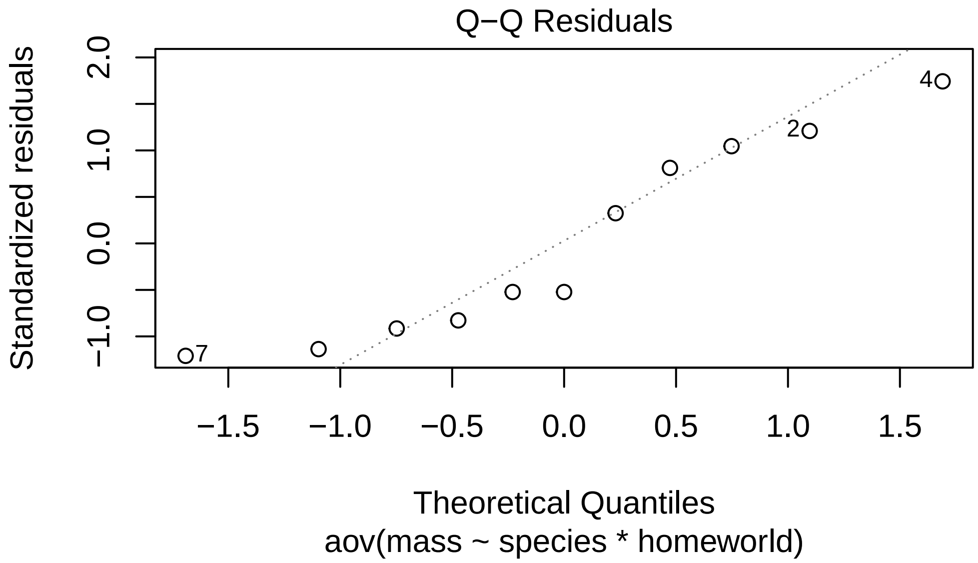
```
group  3  0.3306 0.8037
       8
```

Normallity of Residuals

Graphical

```
plot(aov_factorial,2)                                                    ①
```

① Plot ANOVA residuals using `plot()` function



Statistical

```
fanova_residuals <- aov_factorial$residuals                              ①

shapiro.test(fanova_residuals)                                           ②

psych::describe(fanova_residuals)                                        ③
```

① Create residual data set
② Statistical test of residual normality using `shapiro.test()` function

③ Descriptive statistics of residuals using `describe()` function in the `psych` package

```
    Shapiro-Wilk normality test

data:  fanova_residuals
W = 0.90166, p-value = 0.1666

   vars  n mean    sd median trimmed   mad    min max range skew kurtosis   se
X1    1 12    0 21.44     -6   -1.67 22.61 -23.33  40 63.33 0.43     -1.4 6.19
```