ADA2: Class 09, Ch 05b Paired Experiments and Randomized Block Experiments: Two-way Factor design

[Advanced Data Analysis 2](https://StatAcumen.com/teach/ada12, Stat 428/528, Spring 2023, Prof. Erik Erhardt, UNM

AUTHOR
Sina Mokhtar

PUBLISHED
February 17, 2023

Kangaroos skull measurements: crest width

What effect does sex and species have on the crest width of a kangaroo skull?

The data to be analyzed here are selected skull measurements on 148 kangaroos of known sex and species. There are 11 columns of data, corresponding to the following features. The measurements are in meters/10000 (mm/10).

column	Variable name	Description
1 *	sex	sex (1=M, 2=F)
2 *	species	species (0=M. giganteus, 1=M.f. melanops, 2=M.f. fuliginosus)
3	pow	post orbit width
4	rw	rostal width
5	sopd	supra-occipital - paroccipital depth
6 *	cw	crest width
7	ifl	incisive foramina length
8	ml	mandible length
9	mw	mandible width
10	md	mandible depth
11	arh	ascending ramus height

Some of the observations in the data set are missing (not available). These are represented by a period `.`, which in the `read_csv()` function is specified by the `na = "."` option.

```
library(erikmisc)
```

```
── Attaching packages ──────────────────────────────────── erikmisc 0.1.18 ──
```

```
✓ tibble 3.1.8      ✓ dplyr  1.0.10
```

```
── Conflicts ──────────────────────────────────────── erikmisc_conflicts() ──
✖ dplyr::filter() masks stats::filter()
✖ dplyr::lag()    masks stats::lag()

erikmisc, solving common complex data analysis workflows
  by Dr. Erik Barry Erhardt <erik@StatAcumen.com>
```

```r
library(tidyverse)
```

```
── Attaching packages ─────────────────────────────────────── tidyverse 1.3.2
──

✓ ggplot2 3.4.0     ✓ purrr   1.0.1
✓ tidyr   1.3.0     ✓ stringr 1.5.0
✓ readr   2.1.3     ✓ forcats 1.0.0
── Conflicts ────────────────────────────────────── tidyverse_conflicts() ──
✖ dplyr::filter() masks stats::filter()
✖ dplyr::lag()    masks stats::lag()
```

```r
# First, download the data to your computer,
#   save in the same folder as this Rmd file.

dat_kang <-
  read_csv(
    "ADA2_CL_09_kang.csv"
  , na = c("", ".")
  ) %>%
  # subset only our columns of interest
  select(
    sex, species, cw
  ) %>%
  # make dose a factor variable and label the levels
  mutate(
    sex     = factor(sex    , labels = c("M","F"))
  , species = factor(species, labels = c("Mg", "Mfm", "Mff"))
  )
```

```
Rows: 148 Columns: 11
── Column specification ────────────────────────────────────────────────────────
Delimiter: ","
dbl (11): sex, species, pow, rw, sopd, cw, ifl, ml, mw, md, arh

ℹ Use `spec()` to retrieve the full column specification for this data.
ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
# remove observations with missing values
n_start  <- nrow(dat_kang)
dat_kang <- na.omit(dat_kang)
n_keep   <- nrow(dat_kang)
```

```
n_drop   <- n_start - n_keep
cat("Removed", n_start, "-", n_keep, "=", n_drop, "observations with missing values.")
```

```
Removed 148 - 148 = 0 observations with missing values.
```

```
# The first few observations
head(dat_kang)
```

```
# A tibble: 6 × 3
  sex   species    cw
  <fct> <fct>   <dbl>
1 M     Mg        153
2 M     Mg        141
3 M     Mg        144
4 M     Mg        116
5 M     Mg        120
6 M     Mg        188
```

(1 p) Interpret plots of the data, distributional centers and shapes

The side-by-side boxplots of the data compare the crest widths across the 6 combinations of sex and species. Comment on the distributional shapes and compare the typical crest widths across groups.

```
# Calculate the cell means for each (sex, species) combination
# Group means
kang_mean     <- dat_kang %>%                           summarise(m = mean(cw))
kang_mean_x   <- dat_kang %>% group_by(sex)          %>% summarise(m = mean(cw)) %>% ungroup()
kang_mean_s   <- dat_kang %>% group_by(species)      %>% summarise(m = mean(cw)) %>% ungroup()
kang_mean_xs  <- dat_kang %>% group_by(sex, species) %>% summarise(m = mean(cw)) %>% ungroup()
```

```
`summarise()` has grouped output by 'sex'. You can override using the `.groups`
argument.
```

```
kang_mean
```

```
# A tibble: 1 × 1
      m
  <dbl>
1  123.
```

```
kang_mean_x
```

```
# A tibble: 2 × 2
  sex       m
  <fct> <dbl>
1 M      111.
2 F      136.
```

```
kang_mean_s
```

```
# A tibble: 3 × 2
  species     m
  <fct>   <dbl>
1 Mg       110.
2 Mfm      116.
3 Mff      144.
```

```
kang_mean_xs
```

```
# A tibble: 6 × 3
  sex   species     m
  <fct> <fct>   <dbl>
1 M     Mg       103.
2 M     Mfm      102.
3 M     Mff      128.
4 F     Mg       117.
5 F     Mfm      128.
6 F     Mff      161
```

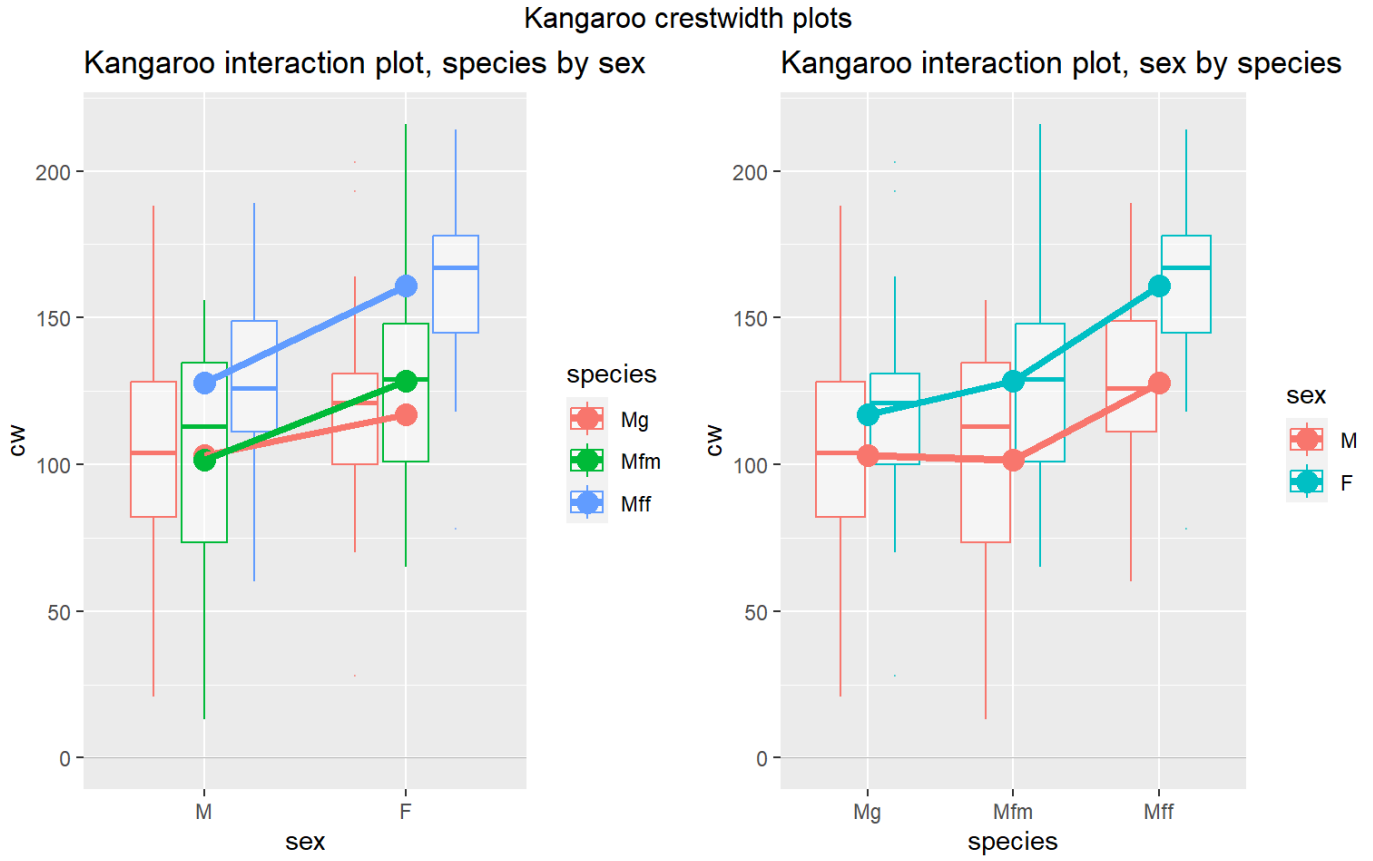
```
# Interaction plots, ggplot
library(ggplot2)
p1 <- ggplot(dat_kang, aes(x = sex, y = cw, colour = species))
p1 <- p1 + geom_hline(aes(yintercept = 0), colour = "black"
                  , linetype = "solid", size = 0.2, alpha = 0.3)
```

```
Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
ℹ Please use `linewidth` instead.
```

```
p1 <- p1 + geom_boxplot(alpha = 0.5, outlier.size=0.1)
p1 <- p1 + geom_point(data = kang_mean_xs, aes(y = m), size = 4)
p1 <- p1 + geom_line(data = kang_mean_xs, aes(y = m, group = species), size = 1.5)
p1 <- p1 + labs(title = "Kangaroo interaction plot, species by sex")
#print(p1)

p2 <- ggplot(dat_kang, aes(x = species, y = cw, colour = sex))
p2 <- p2 + geom_hline(aes(yintercept = 0), colour = "black"
                  , linetype = "solid", size = 0.2, alpha = 0.3)
p2 <- p2 + geom_boxplot(alpha = 0.5, outlier.size=0.1)
p2 <- p2 + geom_point(data = kang_mean_xs, aes(y = m), size = 4)
p2 <- p2 + geom_line(data = kang_mean_xs, aes(y = m, group = sex), size = 1.5)
p2 <- p2 + labs(title = "Kangaroo interaction plot, sex by species")
#print(p2)

library(gridExtra)
```

```
Attaching package: 'gridExtra'

The following object is masked from 'package:dplyr':

    combine
```

```
grid.arrange(grobs = list(p1, p2), nrow=1, top="Kangaroo crestwidth plots")
```



Solution

[answer] The First plot: The distribution between groups in terms of spread is not very different and we have roughly a symmetric distribution for all groups. the means for male sex are lower in compare to female groups. the Mfm and Mg species are about the same but the Mff species are a bit larger.
The second plot: males sex group have smaller cw in compare to female.In the male group, the Mfm and Mg species have the same cw size but cw increase in Mff species. In female group, Mg species have the lowest cw size, the cw size increase a bit in Mfm species and Mff have the largest cw size in compare the other two species. The increase of cw size between Mfm and Mff is parallel so there is no interaction, however there may be interaction in Mg and Mfm species

(1 p) Do the plots above suggest there is an interaction?

Do the lines for each group seem to be very different from parallel?

Solution

[answer] No the lines seems parallel however there may be a bit interaction between Mfm and Mg groups.

Fit the two-way interaction model

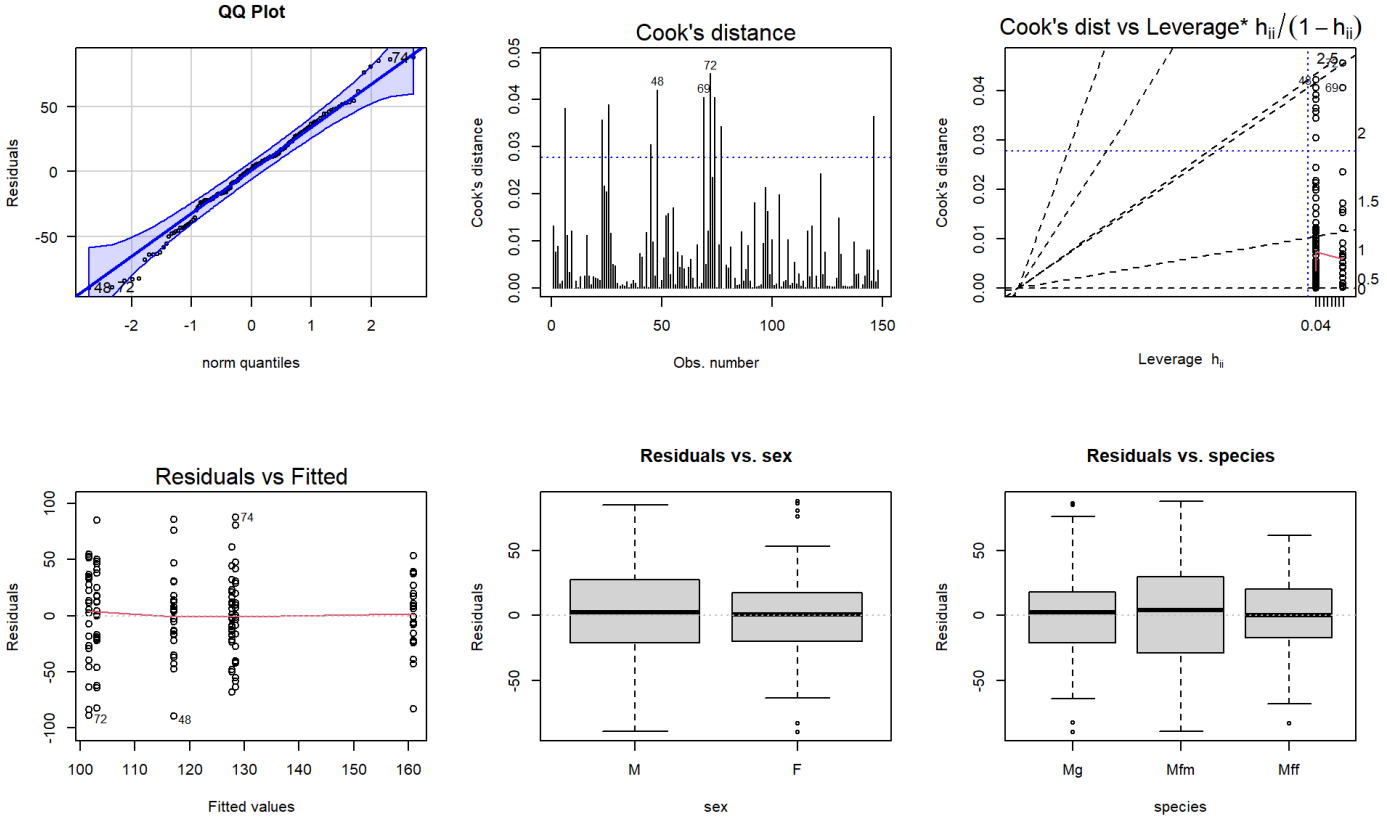
Here it is.

```
lm_cw_x_s_xs <-
  lm(
    cw ~ sex * species
  , data = dat_kang
  , contrasts = list(sex = contr.sum, species = contr.sum)
  )
```

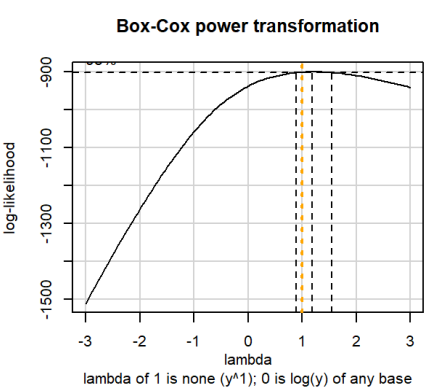
(1 p) Check model assumptions for full model

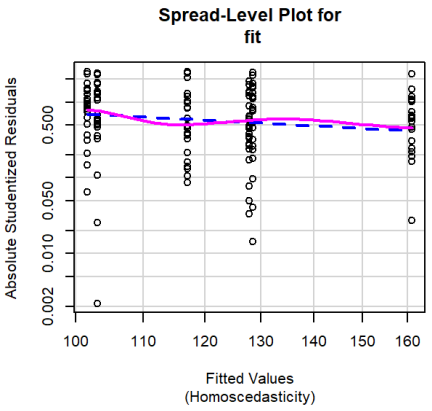
Recall that we assume that the full model is correct before we perform model reduction by backward selection.

```
# plot diagnostics
e_plot_lm_diagostics(lm_cw_x_s_xs)
```

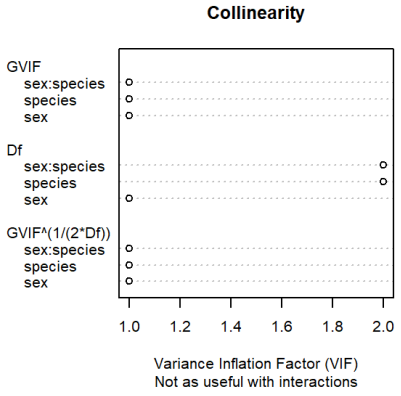


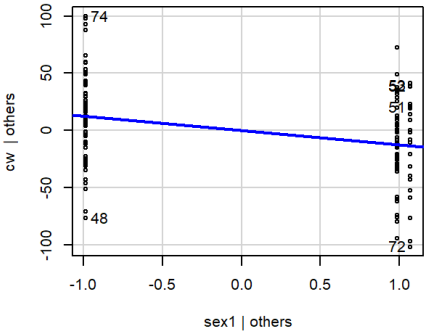
```
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 3.094663, Df = 1, p = 0.078549

there are higher-order terms (interactions) in this model
consider setting type = 'predictor'; see ?vif
```

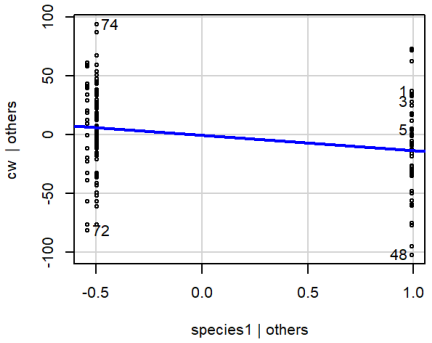


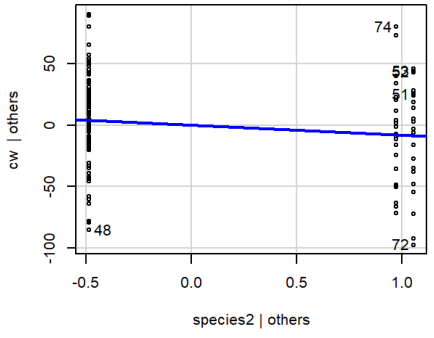


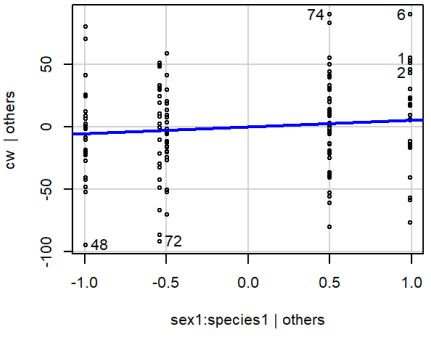

```
Warning in e_plot_lm_diagostics(lm_cw_x_s_xs): Note: Collinearity plot
unreliable for predictors that also have interactions in the model.
```

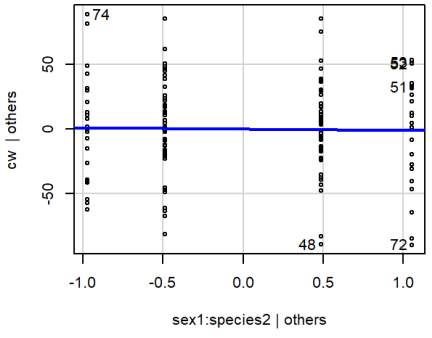












Solution

[answer] The residuals are roughly distributed normal based on QQplot. there are no significant outlier and the variances looks constant in all groups. based on box-cox plot we do not need transformation.

(1 p) ANOVA table, test for interaction

Provide your conclusion for the test for interaction.

```
library(car)
```

```
Loading required package: carData


Attaching package: 'car'

The following object is masked from 'package:purrr':

    some

The following object is masked from 'package:dplyr':

    recode
```

```
Anova(lm_cw_x_s_xs, type=3)
```

```
Anova Table (Type III tests)

Response: cw
             Sum Sq  Df  F value    Pr(>F)
(Intercept) 2244042   1 1643.795 < 2.2e-16 ***
sex           22556   1   16.523 7.927e-05 ***
species       34195   2   12.524 9.788e-06 ***
sex:species    2367   2    0.867    0.4224
Residuals    193853 142
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Solution

[answer] the interaction is not significant based on 95% confidence Interval, however the sex and species are both significant.

(4 p) Reduce to final model, test assumptions

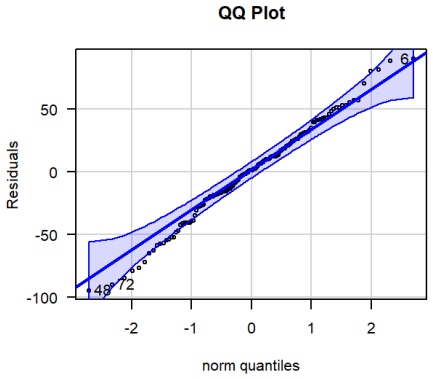
If the model can be simplified (because interaction is not significant), then refit the model with only the main effects. Test whether the main effects are significant, reduce further if sensible. Test model assumptions of your final model.

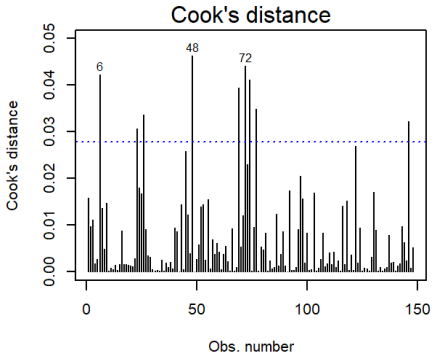
Solution

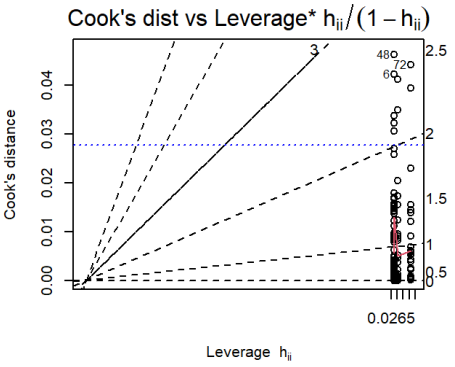
[answer]

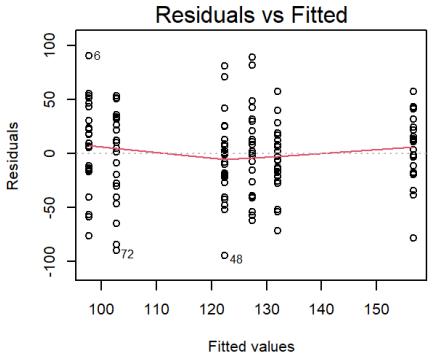
```
lm.reduced = lm(
    cw ~ sex + species
    , data = dat_kang
    , contrasts = list(sex = contr.sum, species = contr.sum)
  )
```

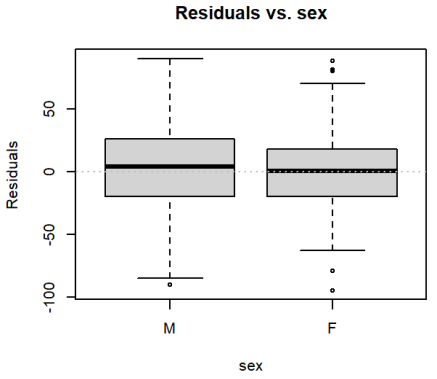
```
# plot diagnostics
e_plot_lm_diagostics(lm.reduced)
```

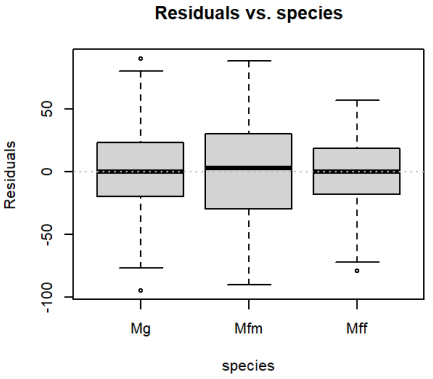




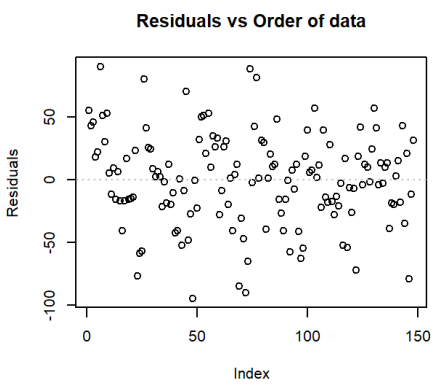


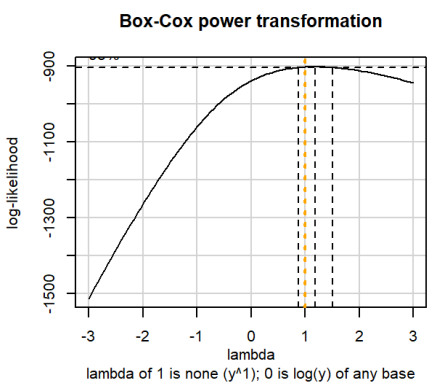


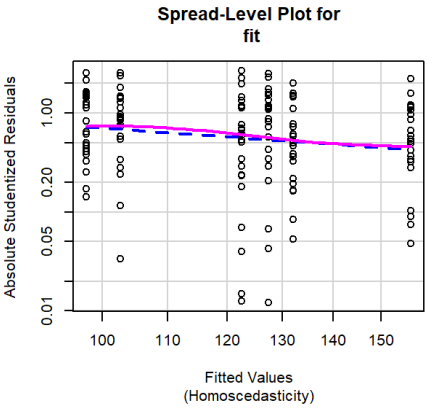




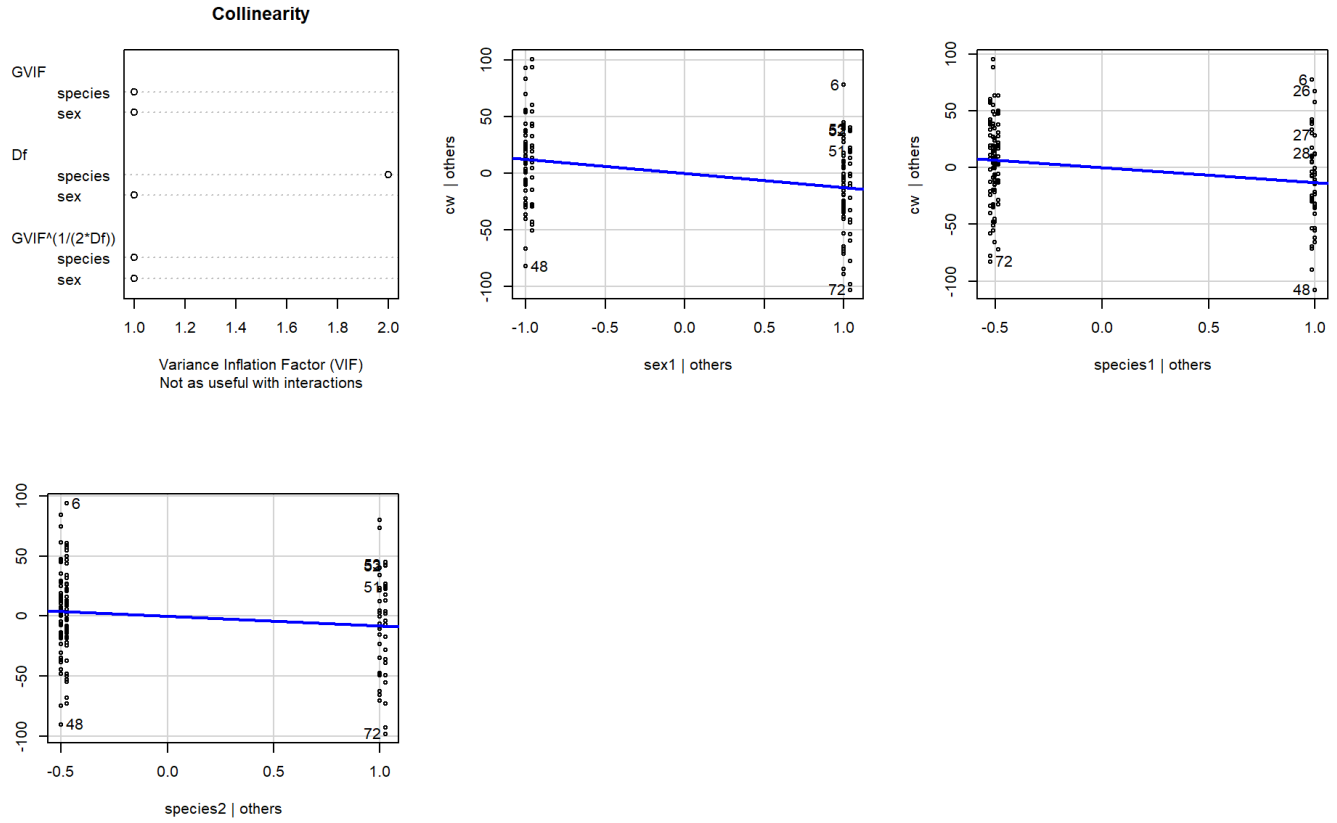

```
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 3.255567, Df = 1, p = 0.071181
```







```
Warning in e_plot_lm_diagostics(lm.reduced): Note: Collinearity plot unreliable
for predictors that also have interactions in the model.
```




```
#hist(lm.reduced$residuals)
```

```
Anova(lm.reduced, type=3)
```

```
Anova Table (Type III tests)

Response: cw
             Sum Sq  Df  F value    Pr(>F)
(Intercept) 2245491   1 1647.901 < 2.2e-16 ***
sex           22511   1   16.520 7.886e-05 ***
species       34163   2   12.536 9.573e-06 ***
Residuals    196220 144
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The residuals are roughly distributed normal based on QQplot (there is a little bit left skewness, but it is not that much severe). there are no significant outlier and the variances looks constant in all groups. based on box-cox plot we do not need transformation. The both main effects are significant.

(2 p) Summarize the differences

Summarize differences, if any, in sexes and species using relevant multiple comparisons. Give clear interpretations of any significant effects.

This code is here to get you started. Determine which comparisons you plan to make and modify the appropriate code. Make the code chunk active by moving the `{R}` to the end of the initial code chunk line.

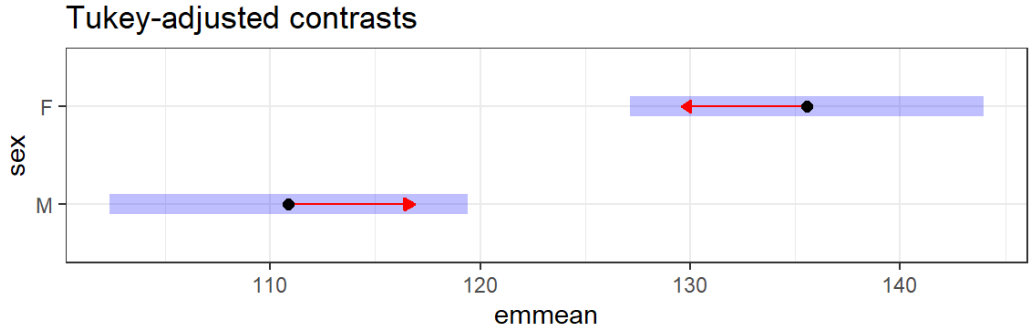
```r
library(emmeans)
# Contrasts to perform pairwise comparisons
cont_kang <- emmeans(lm.reduced, specs = "sex")
# Means and CIs
cont_kang
```

```
 sex emmean   SE  df lower.CL upper.CL
 M      111 4.32 144      102      119
 F      136 4.26 144      127      144

Results are averaged over the levels of: species
Confidence level used: 0.95
```

```r
# Pairwise comparisons
cont_kang %>% pairs()
```

```
 contrast estimate   SE  df t.ratio p.value
 M - F       -24.7 6.07 144  -4.064  0.0001

Results are averaged over the levels of: species
```

```r
# Plot means and contrasts
p <- plot(cont_kang, comparisons = TRUE)
p <- p + labs(title = "Tukey-adjusted contrasts")
p <- p + theme_bw()
print(p)
```



EMM plot interpretation

This **EMM plot (Estimated Marginal Means, aka Least-Squares Means)** is only available when conditioning on one variable. The **blue bars** are confidence intervals for the EMMs; don't ever use confidence intervals for EMMs to perform comparisons – they can be very misleading. The **red arrows** are for the comparisons among means; the degree to which the "comparison arrows" overlap reflects as

much as possible the significance of the comparison of the two estimates. If an arrow from one mean overlaps an arrow from another group, the difference is not significant, based on the adjust setting (which defaults to “tukey”).

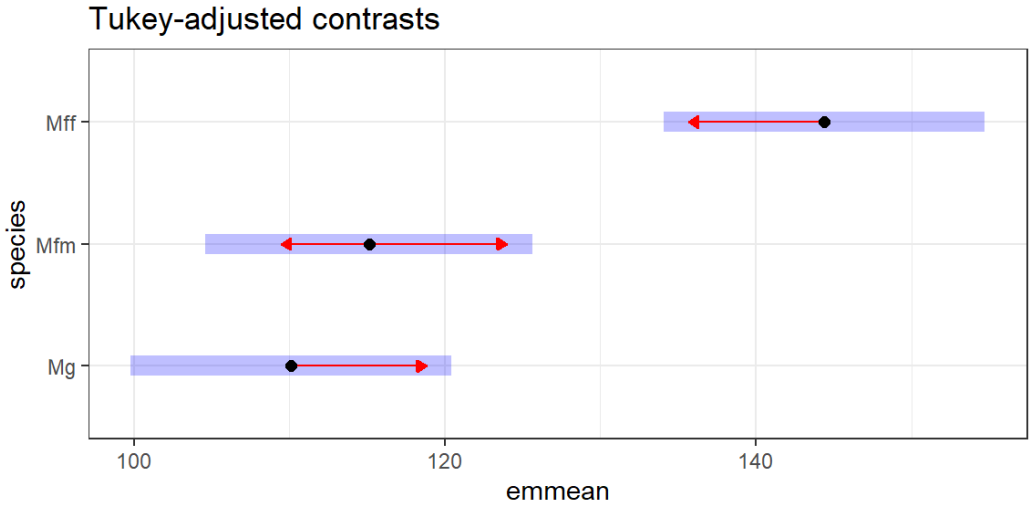
```
cont_kang <- emmeans(lm.reduced, specs = "species")
# Means and CIs
cont_kang
```

```
 species emmean   SE  df lower.CL upper.CL
 Mg         110 5.22 144     99.8      120
 Mfm        115 5.33 144    104.6      126
 Mff        144 5.22 144    134.1      155

Results are averaged over the levels of: sex
Confidence level used: 0.95
```

```
# Pairwise comparisons
cont_kang %>% pairs()
```

```
 contrast  estimate   SE  df t.ratio p.value
 Mg - Mfm     -4.99 7.46 144  -0.669  0.7818
 Mg - Mff    -34.28 7.38 144  -4.643  <.0001
 Mfm - Mff   -29.29 7.46 144  -3.926  0.0004

Results are averaged over the levels of: sex
P value adjustment: tukey method for comparing a family of 3 estimates
```

```
# Plot means and contrasts
p <- plot(cont_kang, comparisons = TRUE)
p <- p + labs(title = "Tukey-adjusted contrasts")
p <- p + theme_bw()
print(p)
```



Solution

[answer]

There are significant size differences between male and female and females in average are 24.7 unit larger than males. there is also significant differences in size between (Mg vs Mff) and (Mfm vs Mff), but there is not significant difference between (Mg vs Mfm) species.

```
Sex       Male    Female
mean      111       136
Groups    -----    -----


Species   Mg    Mfm  Mff
mean      110   115  144
Groups    --------
                    ---
```