ADA2: Class 11, Chs 05 and 07, writing and plotting model equations

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

AUTHOR
Sina Mokhtar

PUBLISHED
February 24, 2023

This assignment is to be printed and hand-written.

In my opinion, some of the most important skills in modeling are:

- writing down a model using indicator variables,
- interpretting model coefficients,
- solving for the predicted value for any combination of predictors, and
- plotting the fitted model.

This assignment applies these skills to two-way factor models (ADA2 Chapter 5) and ANCOVA models with one factor and one continuous predictor (ADA2 Chapter 7).

1. Two-way main-effect model: Kangaroo crest width

Recall these data, results, and the model from Week 05.

```
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>
```

```
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()
```

```
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.
```

```
str(kang)
```

```
tibble [148 × 3] (S3: tbl_df/tbl/data.frame)
 $ sex    : Factor w/ 2 levels "M","F": 1 1 1 1 1 1 1 1 1 1 ...
 $ species: Factor w/ 3 levels "Mg","Mfm","Mff": 1 1 1 1 1 1 1 1 1 1 ...
 $ cw     : num [1:148] 153 141 144 116 120 188 149 128 151 103 ...
`summarise()` has grouped output by 'sex'. You can override using the `.groups`
argument.

# 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

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


Attaching package: 'gridExtra'
```

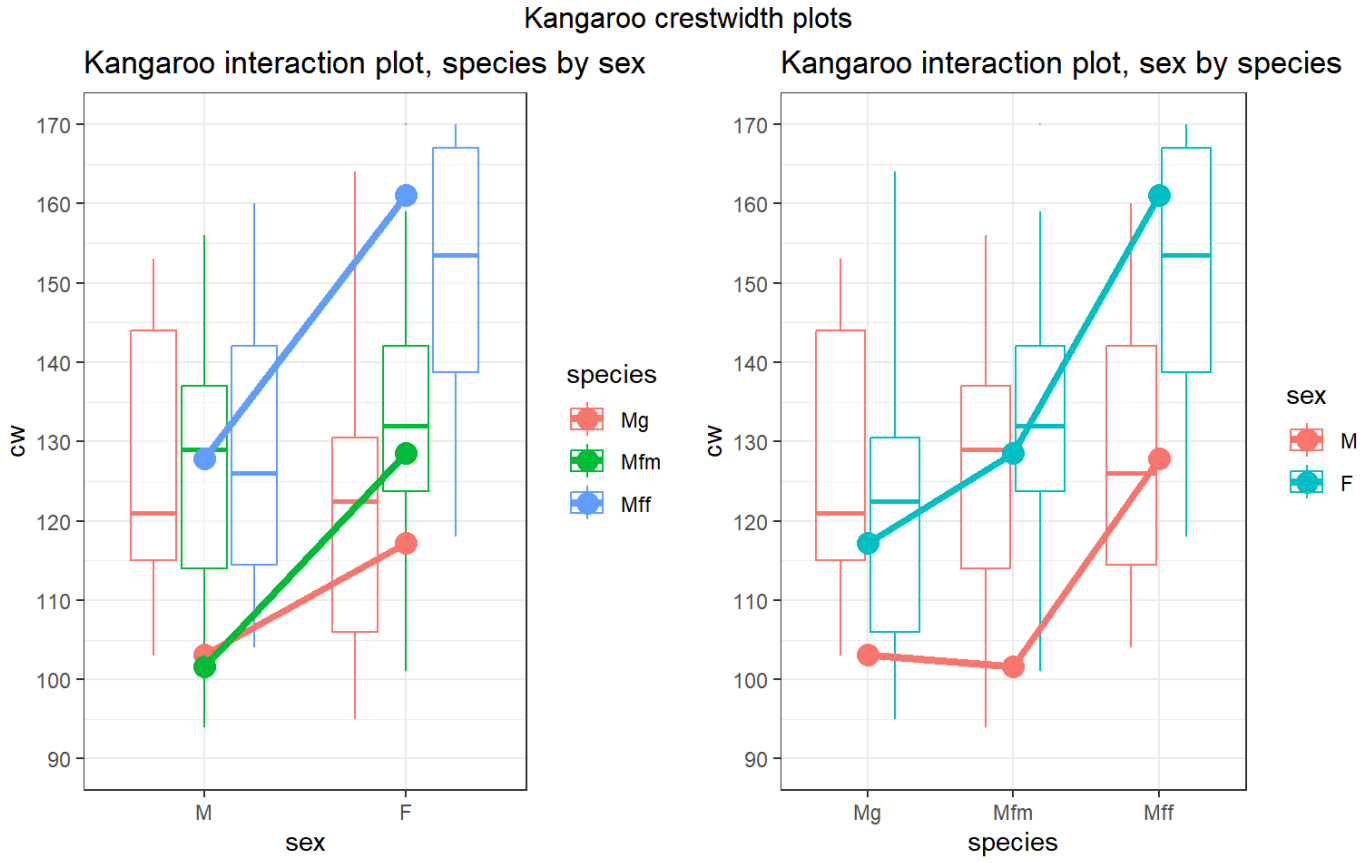
```
The following object is masked from 'package:dplyr':

    combine
Warning: Removed 51 rows containing non-finite values (`stat_boxplot()`).

Warning: Removed 148 rows containing missing values (`geom_hline()`).

Warning: Removed 51 rows containing non-finite values (`stat_boxplot()`).

Warning: Removed 148 rows containing missing values (`geom_hline()`).
```




```r
lm_cw_x_s <-
  lm(
    cw ~ sex + species
  , data = kang
  )
# parameter estimate table
summary(lm_cw_x_s)
```

```
Call:
lm(formula = cw ~ sex + species, data = kang)

Residuals:
    Min      1Q  Median      3Q     Max
```

```
-94.456 -19.746   1.553  23.478  90.216

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   97.784      6.039  16.193  < 2e-16 ***
sexF          24.673      6.070   4.064 7.89e-05 ***
speciesMfm     4.991      7.460   0.669    0.505
speciesMff    34.280      7.383   4.643 7.66e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 36.91 on 144 degrees of freedom
Multiple R-squared:  0.2229,    Adjusted R-squared:  0.2067
F-statistic: 13.77 on 3 and 144 DF,  p-value: 6.057e-08
```

(3 p) Write the fitted model equation.

Use the parameter estimate table above to write out the fitted model equation. Use indicator function notation for categorical variables. First determine what each sex and species number is. The equation looks like: $\hat{y} = [\text{terms}]$.

Solution

$\widehat{cw} = 97.784 + 24.673 * I(sex = F) + 4.991 * I(species = Mfm) + 34.280 * I(species = Mff)$

Intercept represents Sex=M and Species = Mg.

(2 p) Separate model equations.

For each combination of species and sex, write the model.

Solution

Sex	Species	Fitted Model
M	Mg	$\hat{y} = 97.784$
\		
M	Mfm	$\hat{y} = 97.784 + 4.991 = 102.775$
\		
M	Mff	$\hat{y} = 97.784 + 34.280 = 132.064$
\		
F	Mg	$\hat{y} = 97.784 + 24.673 = 122.457$
\		

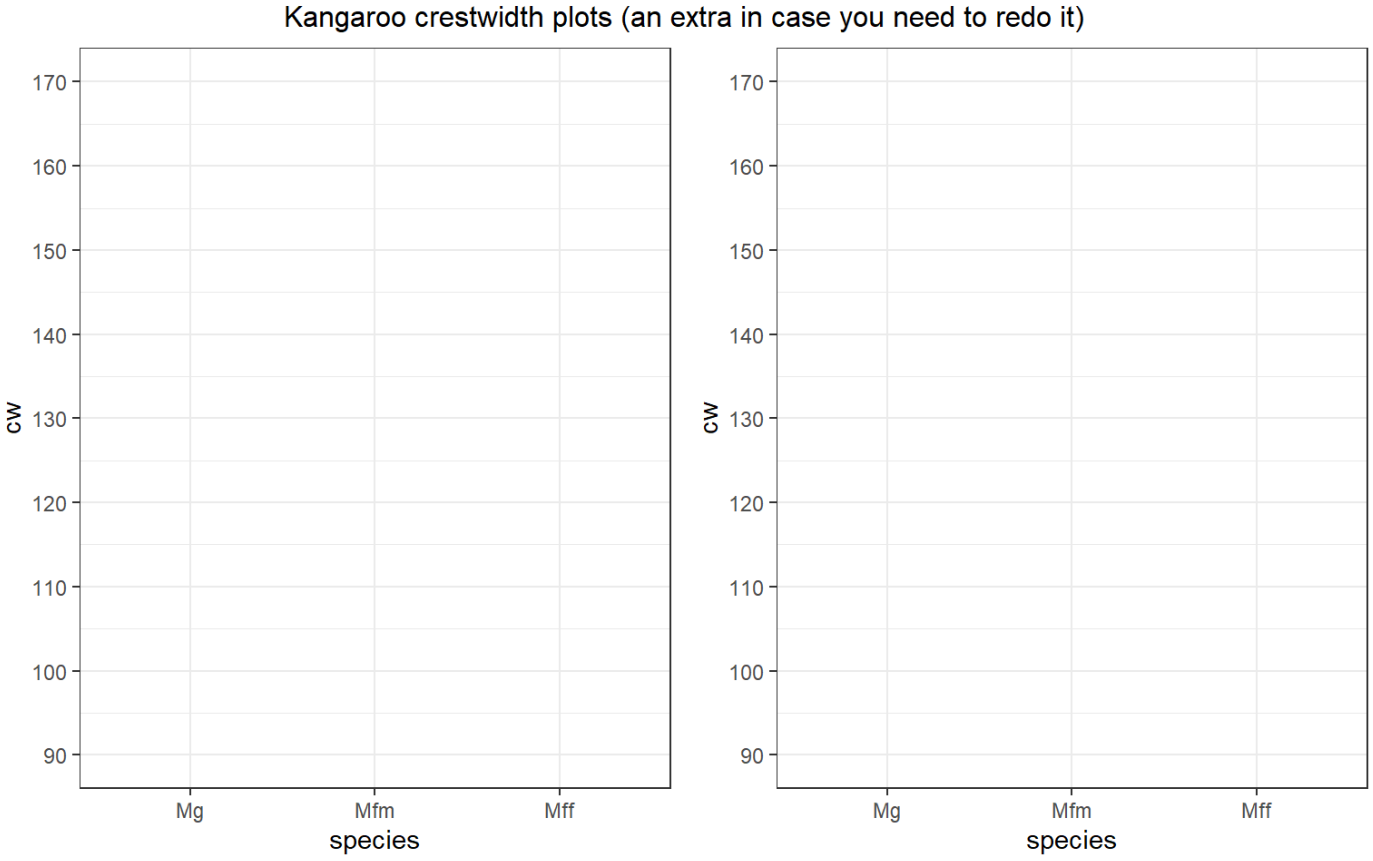
Sex	Species	Fitted Model
F	Mfm	$\hat{y} = 97.784 + 24.673 + 4.991 = 127.448$
\		
F	Mff	$\hat{y} = 97.784 + 24.673 + 34.280 = 156.737$

(0 p) Plot the observed and fitted values.

Use symbols/colors/labels to distinguish between the observed and predicted values and clearly identify the species/sex combinations. Use the minimum about of labeling to make it clear.

Solution

COVID-19 Year, no hand plotting :(



2. ANCOVA model: Faculty political tolerances

A political scientist developed a questionnaire to determine political tolerance scores for a random sample of faculty members at her university. She wanted to compare mean scores adjusted for the age for each of the three categories: full professors (coded 1), associate professors (coded 2), and assistant professors (coded 3). The data are given below. Note the higher the score, the more tolerant the individual.

Below we will fit and interpret a model to assess the dependence of tolerance score on age and rank. (We will assess model fit in a later assignment.)

```
tolerate <-
  read_csv("ADA2_CL_12_tolerate.csv") %>%
  mutate(
    rank = factor(rank)
    # set "3" as baseline level
  , rank = relevel(rank, "3")
  )
```

```
Rows: 30 Columns: 3
── Column specification ────────────────────────────────────────────────────────
Delimiter: ","
dbl (3): score, age, rank

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.
```

```
str(tolerate)
```

```
tibble [30 × 3] (S3: tbl_df/tbl/data.frame)
 $ score: num [1:30] 3.03 4.31 5.09 3.71 5.29 2.7 2.7 4.02 5.52 4.62 ...
 $ age  : num [1:30] 65 47 49 41 40 61 52 45 41 39 ...
 $ rank : Factor w/ 3 levels "3","1","2": 2 2 2 2 2 2 2 2 2 2 ...
```

(3 p) Write the fitted model equation.

Note in the code what the baseline rank is.

```
lm_s_a_r_ar <-
  lm(
    score ~ age * rank
  , data = tolerate
  )
summary(lm_s_a_r_ar)
```

```
Call:
lm(formula = score ~ age * rank, data = tolerate)

Residuals:
     Min       1Q   Median       3Q      Max
-1.34746 -0.28793  0.01405  0.36653  1.07669

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.42706    0.98483   5.511 1.15e-05 ***
```

```
age          -0.01321    0.02948  -0.448   0.6580
rank1         2.78490    1.51591   1.837   0.0786 .
rank2        -1.22343    1.50993  -0.810   0.4258
age:rank1    -0.07247    0.03779  -1.918   0.0671 .
age:rank2     0.03022    0.04165   0.726   0.4751
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6378 on 24 degrees of freedom
Multiple R-squared:  0.5112,    Adjusted R-squared:  0.4093
F-statistic:  5.02 on 5 and 24 DF,  p-value: 0.002748
```

Use the parameter estimate table above to write out the fitted model equation. Use indicator function notation for categorical variables. The equation looks like: $\hat{y} = [\text{terms}]$.

Solution

$ = 5.42706 - 0.01321 * Age + 2.78490 * I(Rank = 1) - 1.22343 * I(Rank = 2)\ - 0.07247 * I(Rank = 1) * Age + 0.03022 * I(Rank = 2) * Age $

Intercept represents Rank = 0(Assistant Prof) and age = 0.

(2 p) Separate model equations.

There's a separate regression line for each faculty rank.

Solution

Rank	Fitted Model
1 FULL	$\hat{y} = (5.42706 + 2.78490) + (-0.01321 - 0.07247) * Age$
\	
2 Assoc	$\hat{y} = (5.42706 - 1.22343) + (0.03022 - 0.01321) * Age$
\	
3 Assis	$\hat{y} = 5.42706 - 0.01321 * Age$

(0 p) Plot the fitted regression lines.

Use symbols/colors/labels to distinguish between the observed and predicted values and clearly identify the rank. Use the minimum about of labeling to make it clear. I recommend plotting each line by evaluating two points then connecting them, for example, by evaluating at age=0 and age=50.

Solution

COVID-19 Year, no hand plotting :(

Faculty tolerance (an extra in case you need to redo it)