Problem Set Week 3

Helge Moes

29-01-2023

Table of Contents

Directions for the student

- Put all R code in code chunks and verbal answers outside code chunks.
- If you cannot make a piece of R code to work, set the code chunk option eval=FALSE.
- Ensure that the R Markdown document knits without problems into a PDF or Word document.
- Submit the R Markdown document and the knitted document on Canvas (under Assignments) before the deadline.

```
#Load all libraries in this code chunk.
library(tidyverse)
```

```
## ── Attaching packages ─────────────────────────────────────── tidyverse
1.3.2 ──
## ✔ ggplot2 3.4.0      ✔ purrr   1.0.1
## ✔ tibble  3.1.8      ✔ dplyr   1.0.10
## ✔ tidyr   1.2.1      ✔ stringr 1.5.0
## ✔ readr   2.1.3      ✔ forcats 0.5.2
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
```

```
library(lubridate)
```

```
## Loading required package: timechange
##
## Attaching package: 'lubridate'
##
## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```
library(broom)
library(papaja)

## Loading required package: tinylabels
```

Questions

A researcher wants to assess the effect of the number of children in the household (famsize) on TV viewing (average number of hours per day) among young children (aged 3-7). In addition, she wants to check whether this effect is different for boys and girls.

Load the data from `tv_viewing.RData`.

1. Apply (two-way) analysis of variance to estimate the effects of the child's sex and family size on TV viewing. Store the results as data objects. Show the ANOVA table.

```
#Add your R code for answering this question here.

# Loading the dataset
load("tv_viewing.RData")
#tv_viewing

# ANOVA to estimate the effects of the child's sex and family size on TV
viewing
tv_viewing_aov <- tv_viewing %>%
  # Filter to include children between 3 and 7
  filter(age >= 3 & age <= 7) %>%
  lm(views ~ sex + famsize, data = ., contrasts= list(sex = contr.sum,
famsize=contr.sum)) %>%
  stats::anova(.)
# ANOVA tibble
tv_viewing_aov

## Analysis of Variance Table
##
## Response: views
##            Df  Sum Sq Mean Sq F value    Pr(>F)
## sex         1   3.880  3.8803 20.4569 6.955e-06 ***
## famsize     3   2.054  0.6847  3.6097   0.01305 *
## Residuals 860 163.127  0.1897
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

#Answer: Based on the ANOVA table, the effect of the child's sex on the views variable is statistically significant. The p-value for the effect of sex is is less than 0.05 (6.955e-06). Therefore, the difference in viewing between girls and boys is unlikely to be due to chance.

The effect of the family size on the views variable is also statistically significant. The p-value for the effect of famsize (0.01305) is less than 0.05.

The F-value for sex is 20.4569 and the F-value for famsize is 3.6097, which indicates the effect of sex on TV viewing is larger than the effect of family size on TV viewing.

Grading	Max points	Awarded
Ex. 1	1	

2. In the following sentence, replace each part between { } (remove the curly brackets) by inline code such that the required information is pulled directly from the original data or or from the results stored for Question 1, when the document is knitted:

"Boys watch significantly more TV (*M* = {average for boys with two decimal places}, *SD* = {standard deviation for boys with two decimal places}) than girls (*M* = {average for girls with two decimal places}, *SD* = {standard deviation for girls with two decimal places}), *F* ({degrees of freedom for main effect sex}, {degrees of freedom residuals}) = {F test value with two decimal places}, $p < 0.001$."

###Answer: "Boys watch significantly more TV (M = round(mean(filter(tv_viewing , sex == round (sd(filter(tv_viewing, sex ==”boy”) %>% select(views)),2)) than girls select(views)),2), SD = round(sd(filter(tv_viewing, sex == “girl”))) %>% select(views)),2)), F (1, 860) = round(tv_viewing_aov$F[1], 2), p < 0.001."

"Boys watch significantly more TV (M = 1.36, SD = 0.44) than girls (M = 1.22, SD = 0.44), F (,) = 20.46, p < 0.001."

```
#Individual checking of variables of Q2:
#mean boy
#round(mean(filter(tv_viewing, sex == "boy")$views),2)
#standard deviation boy
#round(sd(filter(tv_viewing, sex == "boy")$views),2)
#mean girl
#round(mean(filter(tv_viewing, sex == "girl")$views),2)
#standard deviation girl
#round(sd(filter(tv_viewing, sex == "girl")$views),2)
#degrees of freedom for main effect sex
#tv_viewing_aov$Df[1]
#degrees of freedom residuals
#tv_viewing_aov$Df[2]
#F test value with two decimal places
#round(tv_viewing_aov[["F value"]][1],2)
```

The only code should be in the sentence above.

Grading	Max points	Awarded
Ex. 2	2	

3. Display the regression coefficients (from Question 1) as a bar chart with ggplot. All commands, including estimation of the regression model, must be part of one pipe.

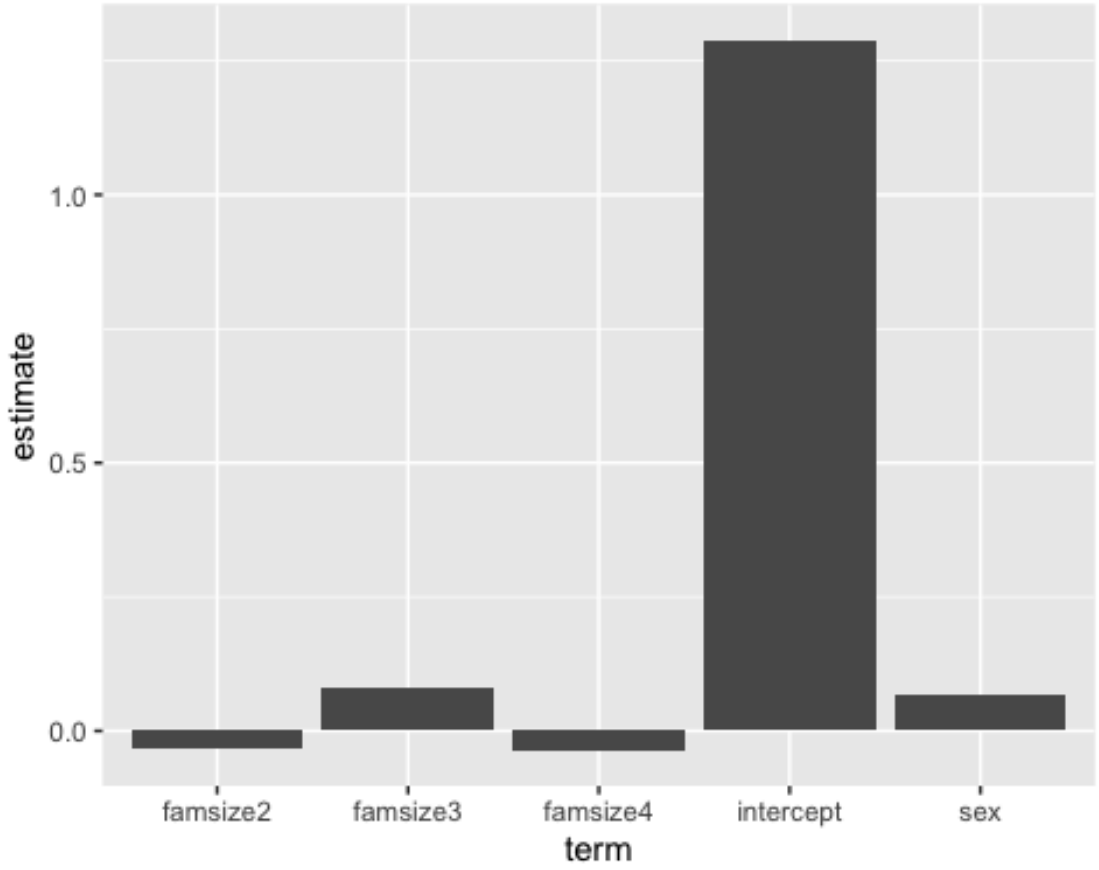
If you decide to retrieve the data directly from the statistical results object, you may have to use `tibble()` or `data_frame()` to change the coefficients into a tibble/data frame.

For a bar chart, this tibble has to contain a variable specifying the coefficient name and a variable with the estimated coefficient value. Both variables can be created from the coefficients vector in the regression data object.

```
#Add your R code for answering this question here.

tv_viewing_coefs<- lm(views ~ sex + famsize, data = tv_viewing, contrasts=
list(sex = contr.sum, famsize=contr.sum))

#Define coefficients and generate tibble
data.frame(term = c("intercept", "sex", "famsize2", "famsize3", "famsize4"),
estimate = coef(tv_viewing_coefs)) %>%
  ggplot(aes(x = term, y = estimate)) +
  geom_bar(stat = "identity")
```



Grading	Max points	Awarded
Ex. 3	2	

4. Correct the function below and apply it to the TV viewing data to show that it works. Note that the variable names must be specified as strings (with quotes) in the function argument if someone uses the function.

```
# Function to calculate the correlation between variables x and y in data
frame data for each category of a factor cat.
```

```
partcorr <- function(data, x, y, cat) {
  corrs <- list()
  for(i in levels(data[,cat])) {
    corrs[i] <- round(cor(data[data[,cat] == i, x], data[data[,cat] == i, y],
method = "pearson"), digits = 2)
  }
  return()
}

partcorr
```

```
## function(data, x, y, cat) {
##   corrs <- list()
##   for(i in levels(data[,cat])) {
##     corrs[i] <- round(cor(data[data[,cat] == i, x], data[data[,cat] == i,
y], method = "pearson"), digits = 2)
##   }
##   return()
## }
```

Answer (part 1):

The function works by adding 'corrs' in the *return* function and adding *(data = tv_viewing, x = "views", y = "age", cat = "sex")* behind the 'partcorr'. This shows that the function works.

```
#Add your R code for answering this question here.

partcorr <- function(data, x, y, cat) {
  corrs <- list()
  for(i in levels(data[,cat])) {
    corrs[i] <- round(cor(data[data[,cat] == i, x], data[data[,cat] == i, y],
method = "pearson"), digits = 2)
  }
  return(corrs)
}

partcorr(data = tv_viewing, x = "views", y = "age", cat = "sex")
```

```
## $boy
## [1] 0.02
##
## $girl
## [1] 0.12
```

Answer (part 2):

I have become intrigued by this function and attempted to understand it. By exploring the function and applying the 'lapply' and 'split()' the function works in a similar fashion. Other data is used in order to examine the functions purpose and use.

```r
#Add your R code for answering this question here.
partcorr2 <- function(data, x, y, cat) {
  corrs <- lapply(split(data, data[,cat]), function(sub){
    cor(sub[[x]], sub[[y]], method = "pearson")
  })
  return(corrs)
}

partcorr2(data=tv_viewing, x="views", y="adhd", cat="sex")
```

```
## $boy
## [1] 0.1263375
##
## $girl
## [1] -0.06196269
```

Grading	Max points	Awarded
Ex. 4	2	

5. Select or create a numeric dependent variable, a numeric predictor, and a categorical predictor from your *Data Project*. Apply a regression model to your project data including an interaction effect. Report the regression results as a publishable (pretty) table that is correctly printed in a PDF or Word document knitted from this R Markdown document. Mind the details: number of digits, p values, lines, caption.

```r
#Add your R code for answering this question here.

Health <- read.csv("Health.csv")

#Numeric dependent variable:
#current_weight

#Numeric predictor:
#sports_per_week

#Categorical predictor:
#healthy_diet

Health_Model <- Health %>%
  #filter invalid variables
  filter(current_weight != 0) %>%
  #adjust the categories
  mutate(diet = fct_collapse(healthy_diet, Unhealthy = c("Unhealthy", "Very
unhealthy", "Below Average"), Average = "Average", Healthy = c("Very
Healthy", "Healthy"))) %>%
  #regression model
  lm(current_weight ~ diet * sports_per_week, data = .) %>%
  apa_print(.)
```

```
## Warning: Unknown levels in `f`: Below Average, Very Healthy

#tibble regression results
apa_table(Health_Model$table)
```

(#tab:unnamed-chunk-7)

**

Predictor	*b*	95% CI	*t*	*df*	*p*
Intercept	153.78	[146.22, 161.33]	40.02	368	< .001
DietBelow average	5.58	[-6.67, 17.82]	0.90	368	.371
DietHealthy	-9.77	[-23.47, 3.93]	-1.40	368	.162
DietUnhealthy	30.40	[16.65, 44.14]	4.35	368	< .001
DietVery healthy	-67.38	[-128.55, -6.21]	-2.17	368	.031
Sports per week	0.14	[-4.19, 4.47]	0.06	368	.949
DietBelow average × Sports per week	4.40	[-4.87, 13.68]	0.93	368	.351
DietHealthy × Sports per week	0.42	[-5.79, 6.64]	0.13	368	.893
DietUnhealthy × Sports per week	-6.66	[-20.57, 7.25]	-0.94	368	.347
DietVery healthy × Sports per week	21.01	[-2.14, 44.16]	1.78	368	.075

Grading	Max points	Awarded
Ex. 5	2	
Flawless knitting	1	
Total	10	