Lesson 5: Do Diets Really Work?

CDT I.M.Smrt

Background

One hundred and eighty generally healthy adults were randomly assigned to one of three diet groups (Vegetarian, Keto, and None) and one of three exercise levels (60 min/day, 30 min/day, and None). After two months, participants were asked how many kilograms they had lost.

```
library(tidyverse)
library(ggResidpanel)

## Load data
dietdat <- read_csv('https://raw.githubusercontent.com/jkstarling/MA376/main/WeightLoss.csv')

glimpse(dietdat)
```

```
## Rows: 180
## Columns: 4
## $ id       <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18~
## $ diet     <chr> "Vegetarian", "None", "None", "Keto", "Vegetarian", "None", "~
## $ exercise <chr> "60 min", "30 min", "60 min", "30 min", "60 min", "30 min", "~
## $ wloss    <dbl> 13.96, -2.12, 6.92, 0.68, 4.68, 0.36, 3.72, 8.92, 4.92, -0.84~
```

Before we get started, we will change the “diet” and “exercise” to categorical variables in R.

```
# dietdat <- dietdat %>% mutate(diet = as.factor(xxx)) %>%
  # mutate(exercise = as.factor(xxx))
```

Step 1: Ask a research question For this exercise the research question is ‘Do diets *really* work?’

Can you come up with a refined research question that more accurately describes what we are trying to do?

Step 2: Design a study and collect data

(1) Is this an observational study or an experiment? How are you deciding?

(2) Identify and classify the response and explanatory variables.

Step 3: Explore the data

(3) Create numerical and graphical summaries of the response variable by the treatment variable. Summarize your observations in context. Hint: For a numerical summary, use pipes and the "summarise" function for the count (n()), mean, and standard deviation; for the graphical summary use a histogram or other appropriate plot.

```
# diet.means <- dietdat %>% ...
# diet.means

# Create a histogram
# dietdat %>% ggplot(aes(x=wloss))+
#   geom_histogram(bins = 10) + facet_wrap(~diet,ncol=1)

# Create a violin plot
# dietdat %>% ggplot(aes(x=wloss,y=diet)) +
```

(4) Fit a model using the `contr.sum` option for the `diet` variable. Write down the model that you are fitting. How does this compare with a model we could create from the `diet.means` dataframe above?

```
# dietdat <- dietdat %>% mutate(diet.cont = xxx)
# contrasts(dietdat$diet.cont) = xxx
# diet.lm <- dietdat %>% lm(wloss~diet.cont, data = .)
# summary(diet.lm)
#
# contrasts(dietdat$diet.cont) # what is this?
```

(5) Verify that μ here is equal to the overall mean from the dataframe `dietdat`. Will this will always be the case with multiple means? Does it make sense to use the "mean of the means"? Why?

```
# mean(xxx)
# mean(xxx)
```

(6) Calculate the total variability in the response variable (SST) and partition the total variability into its component parts (SSE and SSM). Verify that SST equals SSE+SSM. Hint: See LSNs 3 and 4.

```
# SST <- sum(xxxx); SST
# SSE <- sum(xxxx); SSE
# SSM <- sum(xxxx); SSM
#
# SST == SSE + SSM
#
# # Calculate R squared
# rsquared <- xxxx; rsquared
```

(7) What can we conclude at this point?

Step 4: Draw inferences beyond the data

(8) We want to test some hypotheses so that we can make conclusions about the population, not just our sample. What are our hypotheses in words and 'symbols'?

(9) One statistic we can use to test this hypothesis is R^2. Assuming H_0 is true, create 5000 replicates of R^2. Plot the results. Do we have strong evidence against H_0?

```
# M <- 5000
# stat <- rep(NA, M)
# for(i in 1:M){
#   dietdat$diet.shuff <- xxxx
#   sim.lm <- dietdat %>% xxxxx
#   stat[i] <- summary(sim.lm)$xxxxx
# }
# data.frame(stat) %>% ggplot(aes(x=stat)) + xxxx
```

(10) We can also use another statistic, the F-statistic. It is defined as follows:

$$F = \left[\frac{R^2}{1-R^2}\right] \times \left[\frac{n - \#\,groups}{\#\,groups - 1}\right] = \left[\frac{SSM/df\ for\ model}{SSE/df\ for\ error}\right]$$

Calculate the F-statistic for this model and data:

```
## Theory-based F-test
# n <- xxx
# dfmod <- xxx           # We have three groups -> we estimate X group means
# dfred <- xxx           # We have 180 observations-> estimate overall mean & estimate X group means
# Fstat <- xxxx; Fstat
# Fstat <- xxxx; Fstat
#
# 1 - pf(xxxx)
# pf(xxxx)  # Right-tail test (Ha: greater than)
```

(11) Be sure to check that the validity conditions have been satisfied before making any conclusions. Find the validity conditions on p. 80. Hint: you may find "resid_panel" from the package ggResidpanel useful.

```
* Validity condition #1: samples are independent of each other.
```

```
* Validity condition #2: standard deviations of treatment groups are similar (e.g. largest is not more
```

```
* Validity condition #3: The distributions of the groups are approx symmetric (implying the distributio
```

```
# resid_panel(xxxx, plots = c('hist'), bins = 20)
```

(12) There is an easier way to conduct the F test using an R function. (Hint: use the "aov" function.) Do these numbers look familiar? Discuss.

```
## Easier way to do F-test
# diet.aov <- aov(xxxxx)
# summary(diet.aov)
```

(13) Explain to a lay-person the premise for conducting hypothesis testing using the simulation-based approach for this scenario.

Step 5: Formulate conclusions

(14) Based on your analysis so far, summarize the conclusions you would draw from this study. Be sure to address statistical significance, generalizability, and causation. Also be sure to put your comments into the context of this research study.

Step 6: Look back and ahead

(15) Suggest at least one way you would improve this study if you were to carry it, or a follow-up study, out yourself.

(16) Create a sources of variation diagram.

Observed variation in:	Sources of explained variation:	Sources of unexplained variation:
Inclusion criteria:		
Design:		