Chapter 3 homework

Setup

```
library(tidyverse)
library(here)
library(broom)
```

The course datasets live in your project's `data/` folder. Use `here::here()` so file paths work regardless of where you render from.

```
fitness <- readr::read_csv(here::here("data", "fitness.csv"))
```

---

1. Is this a "normal" group (resting pulse)?

The dataset `fitness.csv` contains (among other variables) resting pulse rate (`RSTPULSE`) for a sample of men. A commonly cited "normal" resting pulse rate for men is 72. We want to assess whether this sample looks consistent with that reference value.

(a) Specify MODEL C, MODEL A, and the null hypothesis

Write both a verbal description and a mathematical statement.

- **MODEL C (compact):** predicts the reference value for every case

$$\text{RSTPULSE}_i = 72 + \varepsilon_i$$

- **MODEL A (augmented):** estimates the sample mean (one-parameter model)

$$\text{RSTPULSE}_i = b_0 + \varepsilon_i$$

- **Null hypothesis:** $H_0 : b_0 = 72$ (equivalently, the population mean resting pulse equals 72)

(b) Estimate both models with `lm()`

A convenient way to fit these with `lm()` is to *re-express* the outcome as a deviation from the null value.

Let $Y_i = \text{RSTPULSE}_i - 72$. Then:

- MODEL C becomes $Y_i = 0 + \varepsilon_i$ (0 parameters)
- MODEL A becomes $Y_i = b_0 + \varepsilon_i$ (1 parameter)

```
fitness <- fitness |>
  mutate(rst_dev = RSTPULSE - 72)

model_c <- lm(rst_dev ~ 0, data = fitness)
model_a <- lm(rst_dev ~ 1, data = fitness)

summary(model_c)
```

```
Call:
lm(formula = rst_dev ~ 0, data = fitness)

Residuals:
   Min     1Q Median     3Q    Max
 -32.0  -24.0  -20.0  -13.5    4.0

No Coefficients

Residual standard error: 20 on 31 degrees of freedom
```

```
summary(model_a)
```

```
Call:
lm(formula = rst_dev ~ 1, data = fitness)

Residuals:
    Min      1Q  Median      3Q     Max
-13.742  -5.742  -1.742   4.758  22.258

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   -18.26       1.49  -12.26 3.29e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.294 on 30 degrees of freedom
```

```
broom::tidy(model_a)
```

```
# A tibble: 1 x 5
  term        estimate std.error statistic  p.value
  <chr>          <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)    -18.3      1.49     -12.3 3.29e-13
```

```
broom::glance(model_a)
```

```
# A tibble: 1 x 12
  r.squared adj.r.squared sigma statistic p.value    df logLik   AIC   BIC
      <dbl>         <dbl> <dbl>     <dbl>   <dbl> <dbl>  <dbl> <dbl> <dbl>
1         0             0  8.29        NA      NA    NA  -109.  222.  225.
# i 3 more variables: deviance <dbl>, df.residual <int>, nobs <int>
```

(c) Calculate PRE

Use:

$$\text{PRE} = \frac{\text{SSE}_C - \text{SSE}_A}{\text{SSE}_C}$$

For `lm` objects, you can get SSE (a.k.a. RSS) with `deviance()`.

```r
sse_c <- deviance(model_c)
sse_a <- deviance(model_a)

pre <- (sse_c - sse_a) / sse_c
pre
```

```
[1] 0.8335267
```

(d) Write a tentative summary

In a short paragraph, summarize what you found and what it suggests substantively. (We are not doing a formal test yet—use your judgment.)

The augmented model seems to be a better fit for the data than the compact model, as moving from the compact model to the augmented model reduces the error by about 83%. This suggests that the group is 'abnormal', as the 'normal' resting pulse rate of 72 is not nearly as good a predictor as the group mean resting pulse rate. This seems to suggest that the group's mean resting pulse rate deviates from the 'normal' rate of 72.

---

2. Did running increase pulse rate?

Use the same dataset to assess whether running increased pulse rate. The variable `RUNPULSE` is post-run pulse rate.

Tip: Create a new variable that captures the *change* in pulse rate.

(a) Specify MODEL C, MODEL A, and the null hypothesis

Let $\Delta_i = \text{RUNPULSE}_i - \text{RSTPULSE}_i$.

- **MODEL C (compact):** no average increase

$$\Delta_i = 0 + \varepsilon_i$$

- **MODEL A (augmented):** estimate the average increase

$$\Delta_i = b_0 + \varepsilon_i$$

- **Null hypothesis:** $H_0 : b_0 = 0$

(b) Estimate both models with `lm()`

```r
fitness <- fitness |>
  mutate(pulse_change = RUNPULSE - RSTPULSE)

model_c <- lm(pulse_change ~ 0, data = fitness)
model_a <- lm(pulse_change ~ 1, data = fitness)

summary(model_c)
```

```
Call:
lm(formula = pulse_change ~ 0, data = fitness)

Residuals:
   Min     1Q Median     3Q    Max
  92.0  111.0  116.0  123.5  136.0

No Coefficients

Residual standard error: 116.4 on 31 degrees of freedom
```

```r
summary(model_a)
```

```
Call:
lm(formula = pulse_change ~ 1, data = fitness)

Residuals:
     Min       1Q   Median       3Q      Max
-23.9032  -4.9032   0.0968   7.5968  20.0968

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  115.903      1.966   58.95   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.95 on 30 degrees of freedom
```

```r
broom::tidy(model_a)
```

```
# A tibble: 1 x 5
  term        estimate std.error statistic  p.value
  <chr>          <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)     116.      1.97      59.0 1.40e-32
```

```r
broom::glance(model_a)
```

```
# A tibble: 1 x 12
  r.squared adj.r.squared sigma statistic p.value    df logLik   AIC   BIC
      <dbl>         <dbl> <dbl>     <dbl>   <dbl> <dbl>  <dbl> <dbl> <dbl>
1         0             0  10.9        NA      NA    NA  -118.  239.  242.
# i 3 more variables: deviance <dbl>, df.residual <int>, nobs <int>
```

(c) Calculate PRE

```r
sse_c <- deviance(model_c)
sse_a <- deviance(model_a)

pre <- (sse_c - sse_a) / sse_c
pre
```

```
[1] 0.9914419
```

(d) Write a tentative summary

In a short paragraph, summarize what you found and what it suggests substantively.

The augmented model seems to be a better fit for the data than the compact model, as moving from the compact model to the augmented model reduces the error by about 99%. Given that the compact model suggests that running does not increase pulse rate and the augmented model estimates the average pulse rate increase from running, this suggests that running did in fact increase pulse rate.

---

3. Conceptual practice: write models and hypotheses

For each prompt below:

1. Specify MODEL C, MODEL A, and the null hypothesis.
2. State the number of parameters in MODEL C and MODEL A.
3. State the number of **unused-but-potential parameters** in MODEL A (degrees of freedom), using the course definition.

Do **not** write your models generically as "$Y = ...$". Use the named dependent variable (e.g., "IQ", "PTSD score", etc.). If a prompt implies a *constructed variable*, define it.

(a) IQ

IQ tests are designed to have mean 100 and standard deviation 15. You give 6 friends an online IQ test. Are your friends smarter than average?

- **MODEL C (compact):** predicts the reference value for every case; 0 parameters

$$\text{IQ}_i = 100 + \varepsilon_i$$

- **MODEL A (augmented):** estimates the sample mean; 1 parameter; 5 degrees of freedom

$$\text{IQ}_i = b_0 + \varepsilon_i$$

- **Null hypothesis:** $H_0 : b_0 = 100$ (the population mean IQ equals 100)

(b) PTSD

The army uses a PTSD test; scores above 37 indicate clinical levels of PTSD. A troop of 43 soldiers is tested at the end of deployment. Are these soldiers, on average, suffering from PTSD?

- **MODEL C (compact):** predicts the reference value for every case; 0 parameters

$$\text{PTSD}_i = 37 + \varepsilon_i$$

- **MODEL A (augmented):** estimates the sample mean; 1 parameter; 42 degrees of freedom

$$\text{PTSD}_i = b_0 + \varepsilon_i$$

- **Null hypothesis:** $H_0 : b_0 = 0$ (the population mean PTSD score equals 0)

(c) Chipotle sales

Chipotle wants to know whether sales have rebounded after an E. coli scare. They have sales in 200 markets *before* the scare and *now*. They compute a difference score. Are sales depressed?

-Let $\Delta_i = \text{MKTSAFTER}_i - \text{MKTSBEFORE}_i$ - **MODEL C (compact):** no change; 0 parameters

$$\Delta_i = 0 + \varepsilon_i$$

- **MODEL A (augmented):** estimate the average change; 1 parameter; 199 degrees of freedom

$$\Delta_i = b_0 + \varepsilon_i$$

- **Null hypothesis:** $H_0 : b_0 = 0$ no change

---

4. With your own data

Please choose a variable from the 2024 General Social Survey. Remember to use `drop_na()` in your pipeline to get rid of missing data.

(a) Describe your dataset

Include enough detail that someone else can understand what you have.

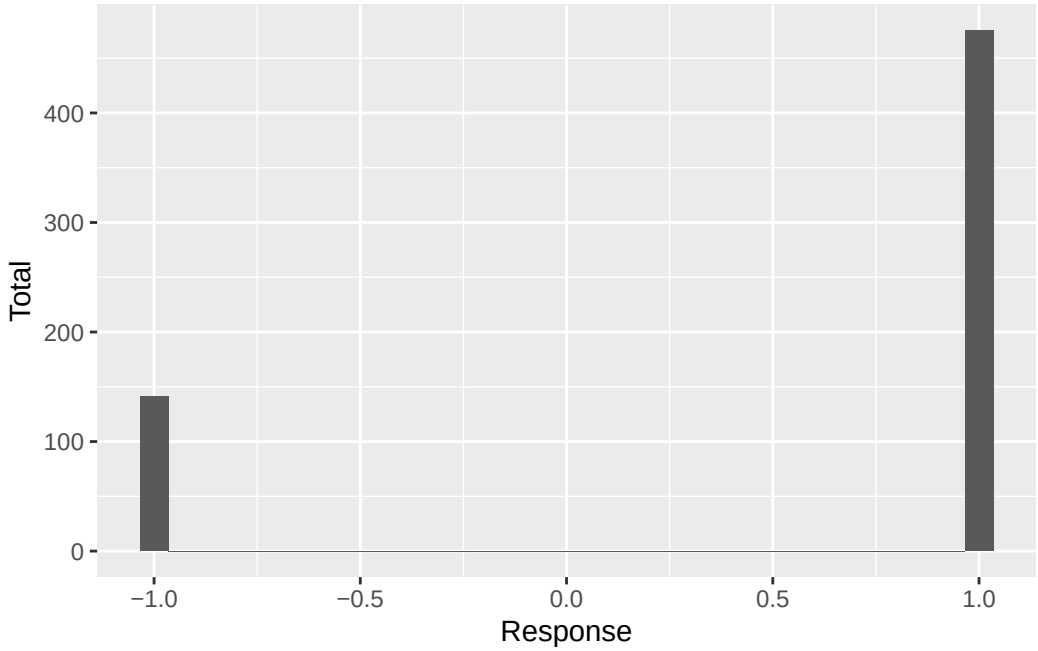
- **(a.1)** What are the units of analysis and how many are there?
- **(a.2)** What is the dependent variable (Y)? How is it measured? What does its distribution look like? (A histogram and/or descriptives are fine.)

```
gss2024 <- readRDS(file = here::here("data", "gss2024.rds"))

gss <- gss2024 |>
  select(courtsnv) |>
  drop_na() |>
  haven::zap_labels()

gss_recoded <- gss |>
  mutate(
    courtsdv = recode(courtsnv,
                      `1` = -1,
                      `2` = 1,
```

```
                          `3` = 0)
    )

gss_recoded |>
  summarize(
    n = n(),
    mean_courts = mean(courtsdv, na.rm = TRUE),
    sd_courts = sd(courtsdv, na.rm = TRUE)
  )
```

```
# A tibble: 1 x 3
      n mean_courts sd_courts
  <int>       <dbl>     <dbl>
1   616       0.542     0.841
```

```
ggplot(data = gss_recoded, aes(x = courtsdv)) +
  geom_histogram() +
  labs(
    x = "Response",
    y = "Total"
  )
```



For these data, the unit of analysis is responses to the 2024 GSS, of which there are 616. The dependent variable is responses to the following question with possible answers of "too harshly", "not harshly enough", or "about right": "In general, do you think the courts in this area deal too harshly or not harshly enough with criminals?" With -1 indicating a response of "too harshly", 0 indicated a response of "about right"*, and 1 indicating a response of "not harshly enough", the mean is .542, indicating that, in general, respondents believe that courts are dealing with criminals "not harshly enough". This is also indicated by the left-skewed histogram representing responses.

*No respondents indicated that courts were dealing with criminals "about right"

(b) Propose a one-parameter question

Think of a question that can be tested with a MODEL C with **0 parameters** and a MODEL A that uses **1 parameter** to estimate central tendency. Write the research question in plain language.

Are respondents in agreement on whether or not courts are dealing with criminals appropriately?

(c) Specify MODEL A, MODEL C, and the null hypothesis

Write both a verbal description and a mathematical statement (use ε_i for error).

- **MODEL C (compact):** predicts that respondents are evenly skewed between 'too harshly' and 'about right' (in disagreement)

$$\text{courtsdv}_i = 0 + \varepsilon_i$$

- **MODEL A (augmented):** estimates the sample mean (one-parameter model)

$$\text{courtsdv}_i = b_0 + \varepsilon_i$$

- **Null hypothesis:** $H_0 : b_0 = 0$ (respondents are evenly distributed between "too harshly" and "not harshly enough")

(d) Estimate both models with `lm()`

```r
gss_recoded <- gss_recoded |>
  mutate(dev = courtsdv - 0)

mod_c <- lm(dev ~ 0,
            data = gss_recoded)

mod_a <- lm(dev ~ 1,
            data = gss_recoded)


summary(mod_c)
```

```
Call:
lm(formula = dev ~ 0, data = gss_recoded)

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

No Coefficients

Residual standard error: 1 on 616 degrees of freedom
```

```r
summary(mod_a)
```

```
Call:
lm(formula = dev ~ 1, data = gss_recoded)

Residuals:
    Min      1Q  Median      3Q     Max
-1.5422  0.4578  0.4578  0.4578  0.4578

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.54221    0.03388      16   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8409 on 615 degrees of freedom
```

(e) Calculate PRE

```
sse_c <- deviance(mod_c)
sse_a <- deviance(mod_a)
pre <- (sse_c - sse_a) / sse_c
pre
```

```
[1] 0.2939893
```

Submission

Render this document to **PDF** and submit the PDF with your code and output.