Chapter 3 homework

Zhe Chen

2026-01-25

Setup

```r
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.

```r
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

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

  In model C, the $RSTPULSE$ of each men is predicted as 72.

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

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

  In Model A, the $RSTPULSE$ of each men is predicted as b_0 (mean of $RSTPULSE$).

- **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)

```r
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
```

```r
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
```

```r
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
```

```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  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.)

- By moving from Model C to Model A, the sum of squared errors reduces by 83.3%, which means Model A predicts much better than Model C. Therefore, it is very likely that we should reject the null hypothesis. Given the estimate of b_0 in Model A is -18.3, the resting pulse rate among men in this sample is lower than 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

$$\Delta_i = 0 + \varepsilon_i$$

  In model C, the Δ of each men is predicted as 0.

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

$$\Delta_i = b_0 + \varepsilon_i$$

  In model A, the Δ of each men is predicted as b_0.

- **Null hypothesis:** $H_0 : b_0 = 0$. In other words, the average change in pulse rate is not zero.

(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
```

```
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
```

```
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

```
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.

- By moving from Model C to Model A, the sum of squared errors reduces by 99.1%, which means Model A predicts much better than Model C. Therefore, it is very likely that we should reject the null hypothesis. Given the estimate of b_0 in Model A is 115.9, the change in pulse rate is higher than 0. In other words, running increased 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 IQ of my friends as 100

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

Number of parameters: 0

- **MODEL A (augmented):** predicts the IQ of my friends as b_0 (mean value)

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

Number of parameters: 1

Unused-but-potential parameters (degrees of freedom): 5

- **Null hypothesis:** $H_0 : b_0 = 100$. The mean value of my friends' IQ is 100, or my friends are smart as average.

(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 PTSD score of soldiers as 37

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

Number of parameters: 0

- **MODEL A (augmented):** predicts the PTSD score of soldiers as b_0 (mean value)

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

Number of parameters: 1

Unused-but-potential parameters (degrees of freedom): 42

- **Null hypothesis:** $H_0 : b_0 = 37$. The mean value of the soldiers' PTSD score is equal to 37, or these soldiers are not suffering from PTSD.

(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?

- **MODEL C (compact):** predicts the difference score as 0.

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

Number of parameters: 0

- **MODEL A (augmented):** predicts the difference score as b_0 (mean value)

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

Number of parameters: 1

Unused-but-potential parameters (degrees of freedom): 199

- **Null hypothesis:** $H_0 : b_0 = 0$. The mean value of the difference score is 0, or the sales are not depressed.

---

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.

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

## select satdemoc (R satisfied with way democracy works in America)
satdemoc <- gss2024 |>
  select(satdemoc) |>
  drop_na() |>
  haven::zap_labels()
glimpse(satdemoc)
```

```
Rows: 3,163
Columns: 1
$ satdemoc <dbl> 2, 1, 3, 4, 4, 3, 2, 2, 4, 2, 1, 2, 4, 3, 4, 3, 3, 3, 2, 3, 3~
```

(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?

The units of analysis are individuals. There are 3163 individuals (the sample size n is 3163).

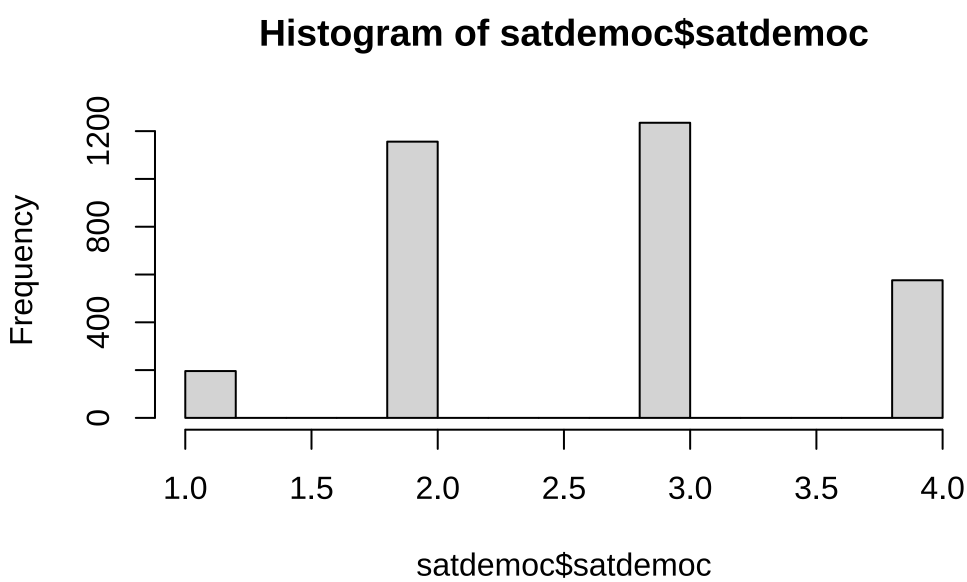
- **(a.2)** What is the dependent variable (Y)? How is it measured? What does its distribution look like? (A histogram and/or descriptives are fine.)

```
# If you have data loaded, you can start with something like:
satdemoc |>
   summarize(
     n = n(),
     mean_y = mean(satdemoc, na.rm = TRUE),
     sd_y = sd(satdemoc, na.rm = TRUE))
```

```
# A tibble: 1 x 3
      n mean_y  sd_y
  <int>  <dbl> <dbl>
1  3163   2.69 0.837
```

The dependent variable Y is satisfaction with way democracy works in America. It is measured in a four-point scale, where 1 = very satisfied, 2 = fairly satisfied, 3 = not very satisfied, 4 = not at all satisfied.

```
hist(satdemoc$satdemoc)
```



In terms of distribution, most observations are at 2 and 3, while fewer are at 1 and 4. In other words, most people prefer mild answers over extreme expressions about their satisfaction with the way democracy works in America.

(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.

Research Question: Whether American people are satisfied with the way democracy works in the United States in general?

(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 the satisfaction score as 2.5 (neutral).

$$\text{Satisfy}_i = 2.5 + \varepsilon_i$$

- **MODEL A (augmented):** predicts the satisfaction score as b_0 (mean value)

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

Null hypothesis: $H_0 : b_0 = 2.5$. The mean value of the satisfaction score is equal to 2.5, or American people are neutral to the way democracy works on average.

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

```r
# Replace Y with your dependent variable and adapt as needed.
# If your null value is mu0, you can use the same deviation trick as in Question 1:
satdemoc <- satdemoc |>
  mutate(satdemoc_dev = satdemoc - 2.5)

sat_model_c <- lm(satdemoc_dev ~ 0, data = satdemoc)
sat_model_a <- lm(satdemoc_dev ~ 1, data = satdemoc)

summary(sat_model_c)
```

```
Call:
lm(formula = satdemoc_dev ~ 0, data = satdemoc)

Residuals:
   Min     1Q Median     3Q    Max 
  -1.5   -0.5    0.5    0.5    1.5 

No Coefficients

Residual standard error: 0.8592 on 3163 degrees of freedom
```

```r
summary(sat_model_a)
```

```

Call:
lm(formula = satdemoc_dev ~ 1, data = satdemoc)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.6927 -0.6927  0.3073  0.3073  1.3073 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.19270    0.01489   12.94   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8374 on 3162 degrees of freedom
```

(e) Calculate PRE

```r
sat_sse_c <- deviance(sat_model_c)
sat_sse_a <- deviance(sat_model_a)
sat_pre <- (sat_sse_c - sat_sse_a) / sat_sse_c
sat_pre
```

```
[1] 0.05030462
```

Submission

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