Chapter 3 Homework

Kaylynn Hiller

2026-01-26

Setup

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.1     v stringr   1.6.0
v ggplot2   4.0.0     v tibble    3.3.1
v lubridate 1.9.4     v tidyr     1.3.2
v purrr     1.2.1
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becom
here() starts at /Users/kaylynnhiller/Desktop/Stats1
```

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

```
Rows: 31 Columns: 9
-- Column specification --------------------------------------------------------
Delimiter: ","
chr (1): _OBSTAT_
dbl (8): AGE, WEIGHT, RUNTIME, RSTPULSE, RUNPULSE, MAXPULSE, OXYGEN, GROUP

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.
```

---

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, meaning that we predict the resting pulse rate for each man should similar to the normal resting heart rate of 72.

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

- **MODEL A (augmented):** estimates the sample mean (one-parameter model). This estimates the average resting pulse rate for men in the sample.

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

- **Null hypothesis:** $H_0 : b_0 = 72$
  - The null hypothesis is that this sample of men will have a resting heart rate of 72 which is the "normal" resting heart rate.

(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)

```
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


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

# 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


# 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()`.

```
[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 potential reduction in error seems large. This would suggest that for our sample of men their resting heart rate is substantially different from the "normal" resting heart rate of 72. Looking at the estimate for Model A (the sample mean), the estimate is negative. This suggests that for our sample, the resting heart rate is lower than the normal resting heart 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 in heart rate after running.

$$\Delta_i = 0 + \varepsilon_i$$

- **MODEL A (augmented):** estimates the average increase in heart rate after running based on our sample.

$$\Delta_i = b_0 + \varepsilon_i$$

- **Null hypothesis:** $H_0 : b_0 = 0$ there is no increase in heart rate after running.

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

```
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
```

```
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
```

```
# 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
```

```
# 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

```
[1] 0.9914419
```

(d) Write a tentative summary

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

There is a very large reduction in error for Model A. This suggests that our change in heart rate from running is very different. The estimate of sample mean for Model A is positive. This suggests that there is an increase in the heart rate of individuals after running. This would make sense based on real world observations of running (my heart rate goes up a lot when running).

---

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 "$Y = \ldots$". 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?

1. Specify MODEL C, MODEL A, and the null hypothesis.

- **MODEL C (compact):** the reference value is 100 for this IQ test. For our sample, the average IQ should be 100.

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

- **MODEL A (augmented):** estimates the sample mean (one-parameter model). This estimates the average IQ for my 6 friends.

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

- **Null hypothesis:** $H_0 : b_0 = 100$, The null hypothesis is that this sample of friends will have an IQ of 100 which is the "normal" IQ.

2. State the number of parameters in MODEL C and MODEL A.

There are 0 parameters in Model C and 1 parameter in Model A.

3. State the number of **unused-but-potential parameters** in MODEL A (degrees of freedom), using the course definition.

There are 5 unused but potential parameters in Model A. One parameter is used to estimate mean. Since there are 6 friends (n = 6), this leaves 5 unused parameters (degrees of freedom).

(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?

1. Specify MODEL C, MODEL A, and the null hypothesis.

- **MODEL C (compact):** the reference value is 37 for this PTSD test.

$$PTSD_i = 37 + \varepsilon_i$$

- **MODEL A (augmented):** estimates the sample mean (one-parameter model). This estimates the average PTSD score for the troop of 43 soldiers.

$$PTSD_i = b_0 + \varepsilon_i$$

- **Null hypothesis:** $H_0 : b_0 = 37$, The null hypothesis is that soldiers will have an average score equal to 37.

2. State the number of parameters in MODEL C and MODEL A.

There are 0 parameters in Model C and 1 parameter in Model A.

3. State the number of **unused-but-potential parameters** in MODEL A (degrees of freedom), using the course definition.

There are 42 unused but potential parameters in Model A. One parameter is used to estimate mean. Since there are 43 soldiers (n = 43), this leave 42 unused parameters (degrees of freedom).

(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?

1. Specify MODEL C, MODEL A, and the null hypothesis.

Let $\Delta_i = \text{SALESNOW}_i - \text{SALESBEFORE}_i$.

- **MODEL C (compact):** the reference value is 0 for analysis. For our sample, their should be no change in sales.

$$\Delta_i = 0 + \varepsilon_i$$

- **MODEL A (augmented):** estimates the sample mean (one-parameter model). This estimates the average difference in sales for the 200 markets.

$$\Delta_i = b_0 + \varepsilon_i$$

- **Null hypothesis:** $H_0 : b_0 = 0$, The null hypothesis is that this sample of markets will have no change in sales before and after an E. coli scare.

2. State the number of parameters in MODEL C and MODEL A.

There are 0 parameters in Model C and 1 parameter in Model A.

3. State the number of **unused-but-potential parameters** in MODEL A (degrees of freedom), using the course definition.

There are 199 unused but potential parameters in Model A. One parameter is used to estimate mean. Since there are 200 markets (n = 200), this leave 199 unused parameters (degrees of freedom).

---

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?

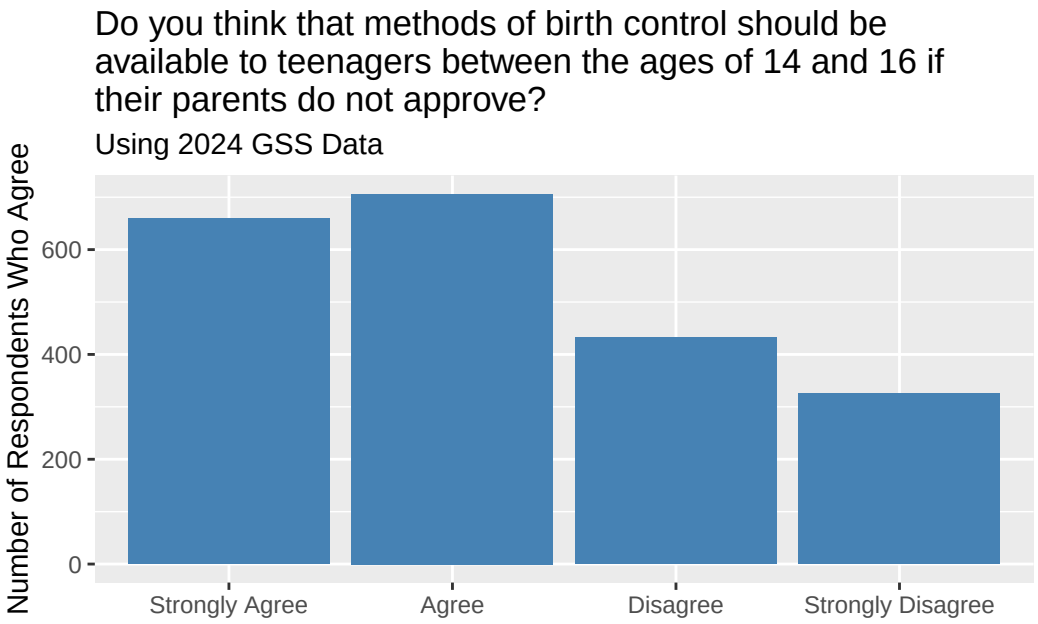
The units of analysis is the individual. Each row in this data is equivalent to the responses of one individual to the survey.

- **(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.)

The dependent variable I chose is pillok. In the GSS, respondents were asked "Do you strongly agree, agree, disagree, or strongly disagree that methods of birth control should be available to teenagers between the ages of 14 and 16 if their parents do not approve?" These responses correlate to the values 1 (strongly agree), 2 (agree), 3 (disagree), 4 (strongly disagree), and NA (no response or not asked).

```
# A tibble: 1 x 5
      n mean_pillok median_pillok var_pillok sd_pillok
  <int>       <dbl>         <dbl>      <dbl>     <dbl>
1  2123        2.20             2       1.09      1.04
```

```
# A tibble: 659 x 1
   pillok
    <dbl>
 1      1
 2      1
 3      1
 4      1
 5      1
 6      1
 7      1
 8      1
 9      1
10      1
# i 649 more rows
```



(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.

On average, do people tend to agree (2) that methods of birth control should be available to teenagers between the ages of 14 and 16 if their parents do not approve?

(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):** the reference value is agree (2) for this analysis.

$$BC_i = 2 + \varepsilon_i$$

- **MODEL A (augmented):** estimates the sample mean (one-parameter model). This estimates the average response for individuals asked this question in the 2024 GSS.

$$BC_i = b_0 + \varepsilon_i$$

- **Null hypothesis:** $H_0 : b0 = 2$, The null hypothesis is that, on average, respondents tend to agree that methods of birth control should be available to teenagers between the ages of 14 and 16 if their parents do not approve.

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

```
Call:
lm(formula = pill_dev ~ 0, data = mydata)

Residuals:
   Min     1Q Median     3Q    Max
    -1     -1      0      1      2

No Coefficients

Residual standard error: 1.061 on 2123 degrees of freedom
```

```
Call:
lm(formula = pill_dev ~ 1, data = mydata)

Residuals:
    Min      1Q  Median      3Q     Max
-1.1997 -1.1997 -0.1997  0.8003  1.8003

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.19972    0.02263   8.825   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.043 on 2122 degrees of freedom
```

(e) Calculate PRE

```
[1] 0.03540141
```

Submission

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