Week 9 Exercises

Exercise 1 First draft

Make sure to upload your "first draft" to Canvas by Sunday night. Also email the PDF to your reviewer below.

Author	**Reviewer**
Caroline	David
Sarah	Taylor
Taylor	Shaka
Bryson	Sarah
David	Bryson
Shaka	Caroline

Exercise 2 Provide feedback

Provide comments on your assigned paper. When providing feedback on colleagues' work, I like the RAP model here. Most importantly, in this model, "focus time and effort on the *highest points of leverage* for improving the document."

The goal is to make suggestions that improve the final product at the end of the semester.

In addition to other comments, please make sure to provide feedback on these three items:

1. Does the paper make a clear *descriptive* claim? It might be helpful to review the key readings on descriptive research here, such as Gerring (2012), Holmes et al. (2024), and de Kadt and Grzymala-Busse (2025).
2. Does the paper make a compelling empirical argument for that descriptive claim?
3. Does the paper clearly explain the substantive/theoretical/normative importance of that descriptive claim? It's important that the defense not rely on any assumed/implied *causal* relationship among the variables.

Exercise 3 Simulate and recover; logit

One important way to test and develop your understandings of statistical models and quantities of interest is to *simulate* a fake data set with known true quantity of interest (e.g., coefficients or first difference) and then *recover* that quantity of interest though your estimation procedure.

Do this for logistic regression. As your quantity of interest, use the first difference as one variable moves from the minimum to the maximum with other variables at their mean or mode.

1. First, simulate a fake data set with a known quantity of interest. The code below handles this.[1]
2. Second, we can use `glm()` and {marginaleffects} to recover the true quantity of interest–it should usually be in the 95% CI.
3. Third, adapt this code to compute the truth for the quantity of interest computed by `avg_comparisons(fit, variables = list(x1 = "minmax"))`.
4. For this particular version of the simulation and recover exercise, the quantities of interest in (2) and (3) seem similar, but they give different answers–the estimate for (3) is much larger than the estimate for (2). Why?

```
# --- simulate a fake data set ---

# set sample size
n <- 100000  # needs to be large enough so that CI isn't super wide
# note: feel free to make the sample size VERY large in this exercise

# create explanatory variables
## numeric variable
x1 <- rnorm(n, 0, 0.5)  # sd = 0.5 is my preferred scale for interpretation
## qualitative variable
x2 <- sample(c("Label 1", "Label 2"),
             size = n,
             replace = TRUE,
             prob = c(0.75, 0.25))  # about 75% will be "Label 1"

# create coefficients
b0 <- 0
b1 <- 1
b2 <- -3

# simulate outcome for normal model
```

[1] In the future, I'll ask you to write this part yourself by adapting this code to other models (e.g., ZINB).

```r
linpred <- b0 + b1*x1 + b2*(x2 == "Label 1")
p <- plogis(linpred)
y <- rbinom(n, size = 1, prob = p)

# combine into data frame
data <- data.frame(y, x1, x2)

# truth: fd as x1 moves from min to max; x2 at mode
pr_hi <- plogis(b0 + b1*max(x1) + b2)
pr_lo <- plogis(b0 + b1*min(x1) + b2)
pr_hi - pr_lo
```

```
[1] 0.323663
```

Solution: (2)

```r
# load packages
library(marginaleffects)

# recover the first quantity of interest
fit <- glm(y ~ x1 + x2, data = data, family = binomial)
comparisons(fit,
            variables = list(x1 = "minmax"),
            newdata = datagrid(x2 = "Label 1"))
```

```
      x2 Estimate Std. Error    z Pr(>|z|)     S 2.5 % 97.5 %
 Label 1    0.331     0.0106 31.2   <0.001 705.6  0.31  0.352

Term: x1
Type: response
Comparison: Max - Min
```

Solution: (3)

```r
# truth: avg fd as x1 moves from min to max; x2 at all observed values
pr_hi <- plogis(b0 + b1*max(x1) + b2*(x2 == "Label 1"))
pr_lo <- plogis(b0 + b1*min(x1) + b2*(x2 == "Label 1"))
mean(pr_hi - pr_lo)
```

```
[1] 0.4504301
```

```
# recover this second quantity of interest
avg_comparisons(fit, variables = list(x1 = "minmax"))
```

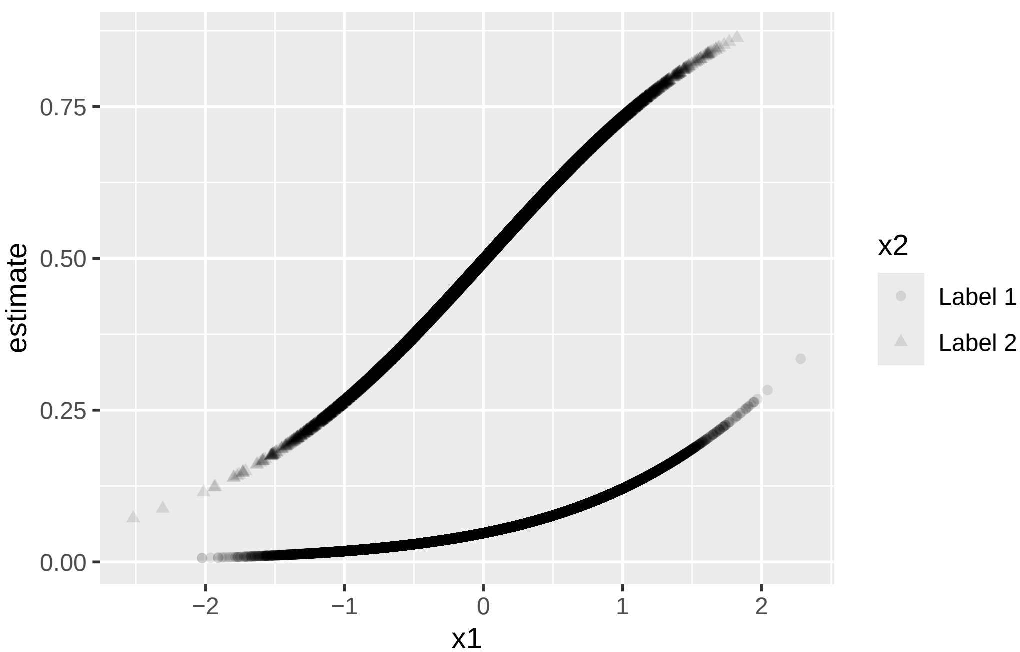
```
 Estimate Std. Error    z Pr(>|z|)   S 2.5 % 97.5 %
    0.457    0.00977 46.8   <0.001 Inf 0.438  0.476

Term: x1
Type: response
Comparison: Max - Min
```

Solution: (3)

The quantity of interest in (2) fixes `x2` to `"Label 1"`. The coefficients `b0`, `b1`, and `b2` make it so that events are *relatively rare* when `x2` equals `"Label 1"`—see the estimated probabilities (i.e., “expected values) below. Because the probabilities are close to zero when `x2` equals `"Label 1"`, `x2` has a smaller effect than when `x2` equals `"Label 2"`. I chose the variables and the coefficients to make this the case.

```
p <- predictions(fit)
ggplot(p, aes(x = x1, y = estimate, shape = x2)) +
  geom_point(alpha = 0.1)
```



The quantity of interest in (3) *averages* across the effects of `x1` for the *observed* values of `x2`. This mean that about 25% of the times, you're getting the big Label 1 effect. And about 75% of the time, you're getting the small Label 2 effect. Averaging these less common big effects and more common small effects gives a larger quantity than looking *only* at the more common small effect.

Exercise 4 Eight schools

In my slides, I motivated hierarchical models with the "309 problem." There's a similar motivational problem called the "eight schools" problem. This example is worth studying carefully.

1. Read this blog post (by Phil Price) on Andrew Gelman's blog: "Everything I need to know about Bayesian statistics, I learned in eight schools."
2. Read the pp. 119-124 of *BDA3* on the eight schools problem.
3. Run the R code here. Note that the notation in my Stan model differs from BDA3.
4. In the model, σ captures the *similarity* of the treatment effects across school districts. I give this parameter a half-Cauchy prior with a scale of 20. (See the note below.) How does changing the scale of the half-Cauchy prior in the Stan model affect the estimates of σ?

   a. Try setting the scale of the half-Cauchy to 5 instead of 20.
   b. Try setting it too 100. Hint: you can use `sigma_draws <- spread_draws(fit, sigma)` to get a convenient data frame of posterior simulations of σ.

Make sure you understand the basic eight schools example and can discuss it in parallel to the 309 example in the slides.

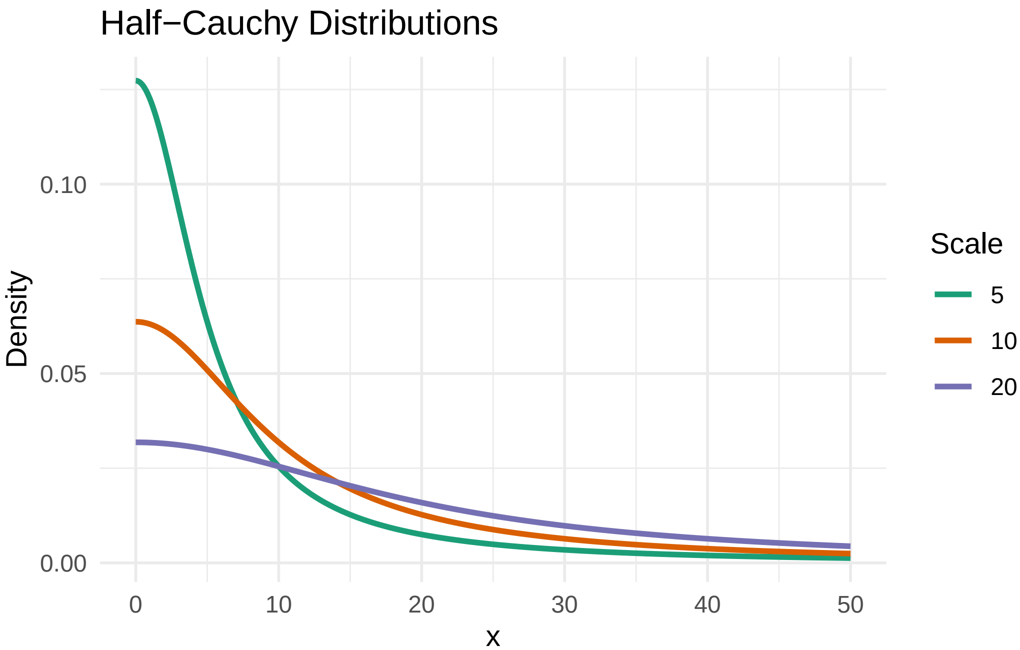
So what's a half-Cauchy?

A Cauchy distribution is a location-scale t distribution with one degree of freedom. The *half-Cauchy* is centered at zero and *folded over*, so that it is just the rightward half. It has only a scale parameter that corresponds to something like the "expected" value of σ.[2] The expected value of the half-Cauchy does not exist, but the median equals the scale parameter. This is a useful *weakly informative* prior for variance parameters that capture similarity in hierarchical models.

In this case, the student-level SAT scores range from 200 to 800, and the treatment effects are somewhere around a 10-point improvement. In the Stan model, σ is the SD around that 10-point effect across schools. $\sigma = 20$ would be huge variation; $sigma = 3$ or $\sigma = 5$ seem "about right." So I think of a half-Cauchy with a scale of 20 as a "weakly informative" prior

[2] The scale parameters of the Cauchy is also denoted as σ, so I just call it the "scale" here.

that only rules out absurdly large values of σ and gently encourages smaller rather than larger values.



Exercise 5 Feelings Toward Donald Trump

The dataset here from the 2016 ANES has data on feeling thermometer ratings of Donald Trump and a `social_group` variable that indicates a race-sex-degree triplet.

```
anes <- read_csv("data/anes-ft-groups.csv") |>
  glimpse()
```

```
Rows: 3579 Columns: 5
-- Column specification --------------------------------------------------------
Delimiter: ","
chr (4): race, sex, college_degree, social_group
dbl (1): ft_donald_trump

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.
```

```
Rows: 3,579
Columns: 5
$ race            <chr> "White", "White", "White", "White", "White", "White", ~
$ sex             <chr> "Male", "Male", "Male", "Male", "Female", "Male", "Mal~
$ college_degree  <chr> "No College Degree", "College Degree", "No College Deg~
$ social_group    <chr> "White; Male; No College Degree", "White; Male; Colleg~
$ ft_donald_trump <dbl> 85, 60, 70, 60, 15, 65, 50, 85, 70, 60, 40, 100, 15, 7~
```

I'm interested in the feelings of each group toward Donald Trump. You can see that the sample averages and sample sizes vary a lot across the groups.[3]

```
library(tinytable)
anes |> group_by(social_group) |>
  summarize(`Number of Respondents` = n(),
            `Sample Average` = mean(ft_donald_trump)) |>
  rename(`Social Group` = social_group) |>
  arrange(desc(`Number of Respondents`)) |>
  tt(digits = 1)
```

Use a hierarchical model to estimate the average for each group.

Solution.

```
library(brms)
f <- ft_donald_trump ~ (1 | social_group)
fit <- brm(f, data = anes, chains = 4, cores = 4)
```

```
library(marginaleffects)
p <- predictions(fit, newdata = datagrid(social_group = unique))
p |>
  select(social_group, estimate, conf.low, conf.high) |>
  arrange(estimate)
```

```
                    social_group Estimate  2.5 % 97.5 %
 Black; Female; College Degree       13.4  5.328   21.5
 Black; Male; College Degree         20.3  8.935   31.6
```

[3]Three of the *possible* groups have zero observations.

Social Group	Number of Respondents	Sample Average
White; Female; No College Degree	784	53
White; Male; No College Degree	675	57
White; Female; College Degree	594	34
White; Male; College Degree	544	42
Other; Female; No College Degree	213	29
Other; Male; No College Degree	213	38
Black; Female; No College Degree	142	21
Other; Male; College Degree	116	36
Black; Male; No College Degree	101	26
Other; Female; College Degree	96	27
Black; Female; College Degree	64	12
Black; Male; College Degree	28	18
White; Other; No College Degree	4	31
Other; Other; No College Degree	3	16
White; Other; College Degree	2	8

```
Black; Female; No College Degree     21.8 16.499   27.3
White; Other; College Degree         26.6  0.226   50.8
Black; Male; No College Degree       26.7 20.464   33.1
Other; Other; No College Degree      26.9  2.828   48.2
Other; Female; College Degree        27.1 20.171   33.5
Other; Female; No College Degree     29.5 25.143   33.8
White; Other; No College Degree      32.0 10.128   54.1
White; Female; College Degree        33.8 31.192   36.5
Other; Male; College Degree          35.8 30.050   41.8
Other; Male; No College Degree       38.3 34.091   42.6
White; Male; College Degree          42.3 39.668   45.2
White; Female; No College Degree     52.4 50.093   54.7
White; Male; No College Degree       56.4 53.996   58.9
```