Name: Yuxuan Zhang

Date: September 24, 2024

HW3

1. (a) By the Bayes rule, we know that

$$p(\theta_j|\bar{y}_j) = \frac{p(\bar{y}_j|\theta_j)p(\theta_j)}{p(\bar{y}_j)}.$$

We also know that the prior distribution is

$$p(\theta_j) = \frac{1}{\sqrt{2\pi\tau^2}} \exp\left(-\frac{(\theta_j - \mu)^2}{2\tau^2}\right),$$

and the likelihood is given by the normal distribution

$$p(\bar{y}_j|\theta_j) = \frac{1}{\sqrt{2\pi\frac{\sigma^2}{n_j}}} \exp\left(-\frac{(\bar{y}_j - \theta_j)^2}{2\frac{\sigma^2}{n_j}}\right).$$

Then, the conditional distribution is

$$\begin{aligned} p(\theta_j|\bar{y}_j) &\propto p(\theta_j)p(\bar{y}_j|\theta_j) \\ &\propto \exp\left(-\frac{(\theta_j - \mu)^2}{2\tau^2} - \frac{(\bar{y}_j - \theta_j)^2}{2\frac{\sigma^2}{n_j}}\right) \\ &= \exp\left(-\frac{1}{2}\left[\frac{n_j}{\sigma^2}(\theta_j - \bar{y}_j)^2 + \frac{1}{\tau^2}(\theta_j - \mu)^2\right]\right). \end{aligned}$$

We then complete the squares:

$$\begin{aligned} \frac{\sigma^2}{n_j}(\theta_j - \mu)^2 + \frac{1}{\tau^2}(\bar{y}_j - \theta_j)^2 &= \frac{n_j}{\sigma^2}(\theta_j^2 - 2\bar{y}_j\theta_j + \bar{y}_j^2) + \frac{1}{\tau^2}(\theta_j^2 - 2\theta_j\mu + \mu^2) \\ &= \left(\frac{n_j}{\sigma^2} + \frac{1}{\tau^2}\right)\theta_j^2 - 2\theta_j(\frac{n_j\bar{y}_j}{\sigma^2} + \frac{\mu}{\tau^2}) + \text{constant} \\ &= \left(\frac{n_j}{\sigma^2} + \frac{1}{\tau^2}\right)\left(\theta_j - \frac{\frac{n_j}{\sigma^2}\bar{y}_j + \frac{1}{\tau^2}\mu}{\frac{n_j}{\sigma^2} + \frac{1}{\tau^2}}\right)^2 + \text{constant}. \end{aligned}$$

Therefore,

$$p(\theta_j|\bar{y}_j) \sim \text{Normal}\left(\frac{\frac{n_j}{\sigma^2}\bar{y}_j + \frac{1}{\tau^2}\mu}{\frac{n_j}{\sigma^2} + \frac{1}{\tau^2}}, \frac{1}{\frac{n_j}{\sigma^2} + \frac{1}{\tau^2}}\right).$$

(b) From part (a), we know that $\mathbb{E}[\theta_j|\bar{y}_j] = \frac{\frac{n_j}{\sigma^2}\bar{y}_j + \frac{1}{\tau^2}\mu}{\frac{n_j}{\sigma^2} + \frac{1}{\tau^2}}$, and $SD[\theta_j|\bar{y}_j] = \sqrt{\frac{1}{\frac{n_j}{\sigma^2} + \frac{1}{\tau^2}}}$. Let $Z = \frac{\theta_j - \mathbb{E}[\theta_j|\bar{y}_j]}{SD[\theta_j|\bar{y}_j]}$. We know that $Z \sim \text{Normal}(0, 1)$. Substitute this into the probability we are solving for, we get

$$\begin{aligned} Pr(\theta_j \in \mathbb{E}[\theta_j|\bar{y}_j] \pm z_{1-\alpha/2}SD[\theta_j|\bar{y}_j]) &= Pr(\mathbb{E}[\theta_j|\bar{y}_j] - z_{1-\alpha/2}SD[\theta_j|\bar{y}_j] \le \theta_j \le \mathbb{E}[\theta_j|\bar{y}_j] + z_{1-\alpha/2}SD[\theta_j|\bar{y}_j]) \\ &= Pr(-z_{1-\alpha/2}SD[\theta_j|\bar{y}_j] \le \theta_j - \mathbb{E}[\theta_j|\bar{y}_j] \le z_{1-\alpha/2}SD[\theta_j|\bar{y}_j]) \\ &= Pr(-z_{1-\alpha/2} \le \frac{\theta_j - \mathbb{E}[\theta_j|\bar{y}_j]}{SD[\theta_j|\bar{y}_j]} \le z_{1-\alpha/2}) \\ &= Pr(-z_{1-\alpha/2} \le Z \le z_{1-\alpha/2}) \\ &= 1 - \alpha. \end{aligned}$$

(c) The width of the usual z-interval is

$$\bar{y}_j \pm z_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n_j}}.$$

The width of the interval from part (b), which is based on the posterior distribution of θ_j, is:

$$\mathbb{E}[\theta_j|\bar{y}_j] \pm z_{1-\alpha/2} \cdot \sqrt{\frac{\sigma^2\tau^2}{n_j\tau^2+\sigma^2}}.$$

Notice that $\frac{\sigma}{\sqrt{n_j}} = \sqrt{\frac{\sigma^2(\tau^2+1/n_j\sigma^2)}{n_j(\tau^2+1/n_j\sigma^2)}}$, and $\sqrt{\frac{\sigma^2\tau^2}{n_j\tau^2+\sigma^2}} = \sqrt{\frac{\sigma^2\tau^2}{n_j(\tau^2+1/n_j\sigma^2)}}$. Since $\tau^2, \sigma^2, n_j > 0$, $\tau^2 + 1/n\sigma^2 > 0$. Thus, $\frac{\sigma}{\sqrt{n_j}} > \sqrt{\frac{\sigma^2\tau^2}{n_j\tau^2+\sigma^2}}$. This shows that the posterior variance is always smaller than $\frac{\sigma^2}{n_j}$, which means that the interval from part (b) is always narrower than that of the usual z-interval. We know that the shrinkage weight is

$$w = \frac{1/\tau^2}{1/\tau^2 + \frac{n_j}{\sigma^2}}.$$

Hence, the width of the interval from part (b) can also be expressed as

$$2z_{1-\alpha/2} \cdot \sqrt{\frac{1-w}{n_j/\sigma^2}}.$$

We know that the weight controls how much the posterior mean is pulled towards μ. As n_j grows, w decreases and the posterior mean relies more on the data $\bar{y}_j$. The posterior interval is narrower because it reflects this additional information from the prior distribution, leading to a more precise estimation of θ_j.

HW3

2024-09-22

```
cd4 = read.table("C:/Users/LEGION/OneDrive - Duke University/610/cd4.dat", header = TRUE)
nels = dget("C:/Users/LEGION/OneDrive - Duke University/610/nels_math_ses.txt")
```

2(a)

```
model1 <- lm(mathscore ~ as.factor(school), data = nels)
z.nels<-abs(model1$res)
leveneTest <- anova(lm(z.nels~as.factor(school), data = nels))
leveneTest
```

```
## Analysis of Variance Table
##
## Response: z.nels
##                     Df Sum Sq Mean Sq F value  Pr(>F)
## as.factor(school)   99   3520  35.557   1.214 0.07887 .
## Residuals         1893  55443  29.288
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Our null hypothesis is that the within-group residual variance across schools are the same. From the result of the Levene's test, we see that the p-value associated with the school factor is 0.07887, greater than 0.05. Therefore, we fail to reject the null hypothesis. There is insufficient evidence of different residual variances across schools, meaning the residual variances are fairly equal across schools in this model.

2(b)

```
model2 = lm(mathscore ~ ses:as.factor(school) + as.factor(school), data = nels)
z.nels<-abs(model2$res)
leveneTest <- anova(lm(z.nels~as.factor(school), data = nels))
leveneTest
```

```
## Analysis of Variance Table
##
## Response: z.nels
##                     Df Sum Sq Mean Sq F value Pr(>F)
## as.factor(school)   99   2938  29.676  1.1404 0.1679
## Residuals         1893  49258  26.021
```

Our null hypothesis is that the within-group residual variance are the same across schools. From the result, we see a p-value of 0.1679, greater than the significance level of 0.05. Therefore, we fail to reject the null hypothesis. This suggest that there is insufficient evidence to conclude that the within-group residual variance are different, meaning that the residual variances are fairly equal across schools in this model as well.

2(c)

Comparing the results of the two models, we see that the mean squared error decreases after adding the 'SES' factor into our model. We observe that the F-statistic decreases, and the p-value increases, suggesting that the additional SES variable is accounting for part of the variation that was previously explained by the group effects alone in Model 1. In model 1, the within-group variances were primarily attributed to differences between schools. However, in model 2, by introducing SES as a covariate, some of the within-group variance is now explained by SES, leading to a lower mean squared residual.

3(a)

```
fit0 <- lm(cd4 ~ time, data = cd4)
fit1 <- lm(cd4 ~ time * trt, data = cd4)
anova(fit0, fit1)
```

```
## Analysis of Variance Table
##
## Model 1: cd4 ~ time
## Model 2: cd4 ~ time * trt
##   Res.Df  RSS Df Sum of Sq      F   Pr(>F)
## 1   1070 2614
## 2   1068 2582  2    32.015 6.6213 0.001387 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the result, we see a F-value of 6.6213, and a p-value of 0.001387, which is lower than the significance level 0.05. This suggests that the model with the treatment effect (fit 0) explains significantly more variation in the CD4 cell percentages than the model without the treatment effect (fit 0). Therefore, we can conclude that there is strong evidence for a treatment effect in the CD4 cell percentages.

3(b)

```
fit2 <- lm(cd4 ~ time * as.factor(pid), data = cd4)
fit2b <- lm(cd4 ~ time * as.factor(pid) + trt * time, data = cd4)
anova(fit2, fit2b)
```

```
## Analysis of Variance Table
##
## Model 1: cd4 ~ time * as.factor(pid)
## Model 2: cd4 ~ time * as.factor(pid) + trt * time
##   Res.Df    RSS Df Sum of Sq F Pr(>F)
## 1    598 300.79
## 2    598 300.79  0         0
```

No, this approach does not adequately evaluate the effects of trt while accounting for across-subject heterogeneity. We found out that the pid and treatment are confounding variables since each pid is only associated with exactly one type of treatment, as shown below.

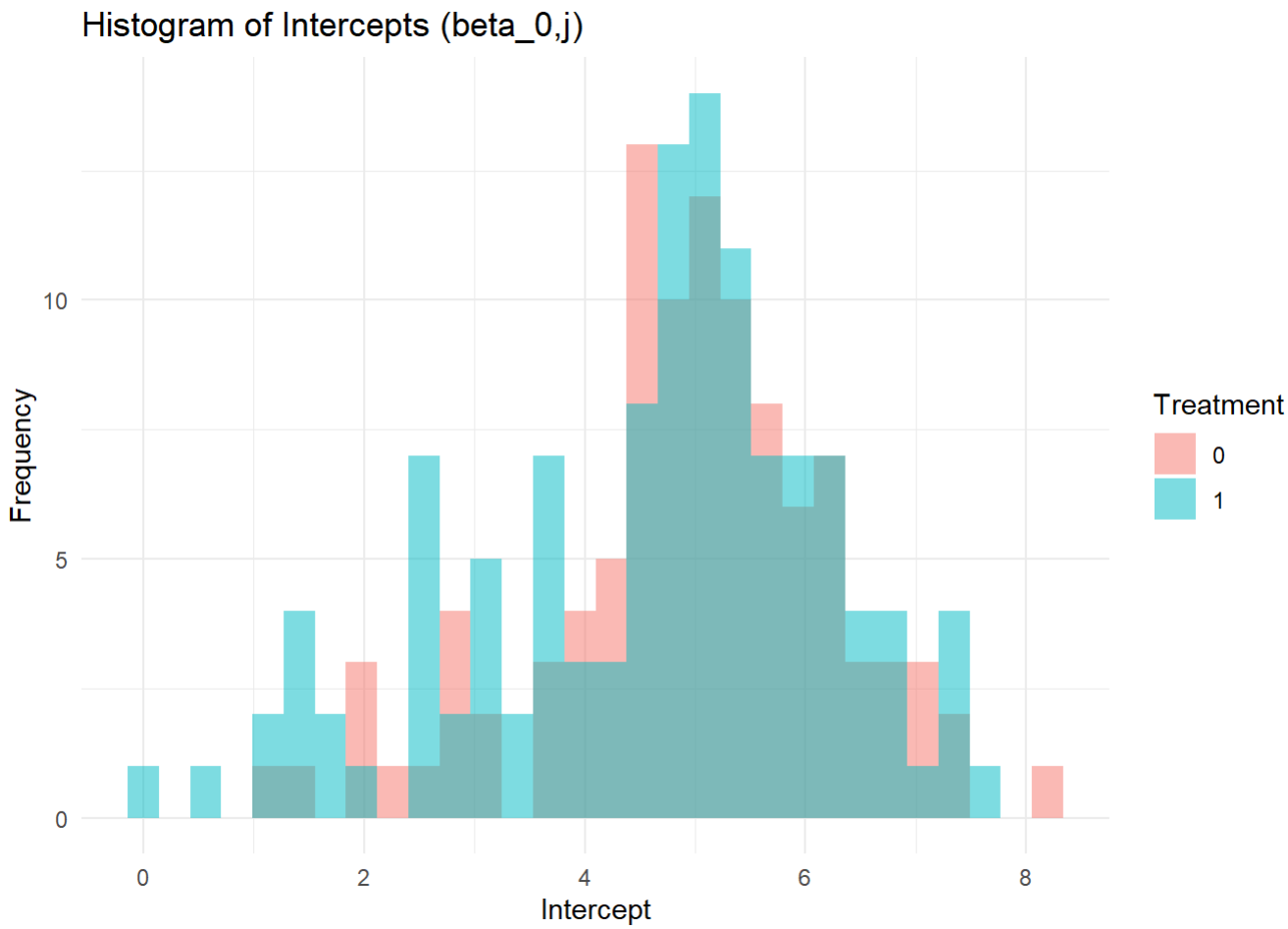
```
library(dplyr)
pid_treatment_check <- cd4 %>%
  group_by(pid) %>%
  summarise(unique_trt = n_distinct(trt))
any(pid_treatment_check$unique_trt > 1)
```

```
## [1] FALSE
```

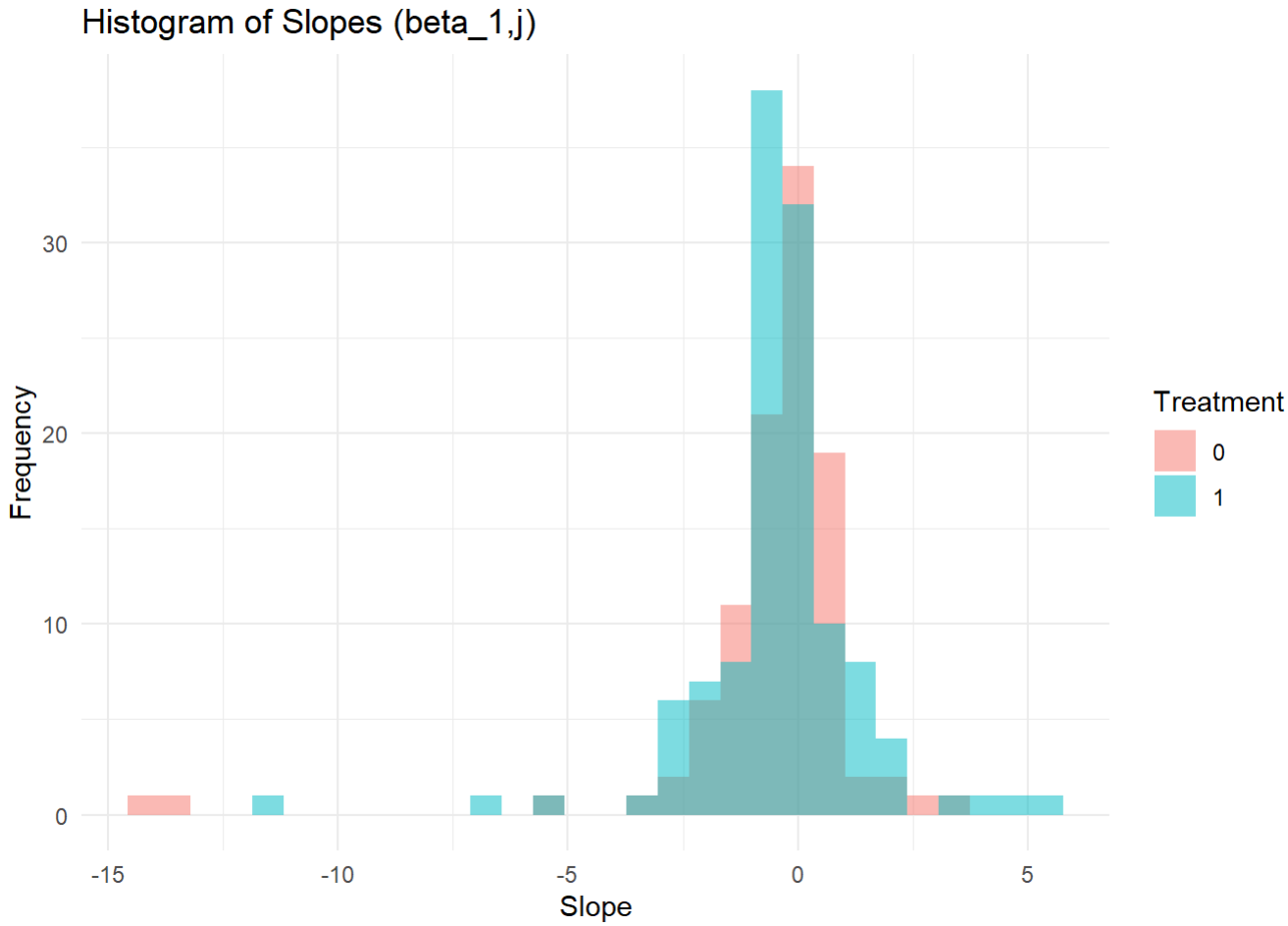
We can also see that the two models are identical from the ANOVA result. When we attempt to separate the effects of treatment from the individual effects of pid, since pid and t rt are confounded, we cannot distinguish the effects between the two with model 2b. Thus, this approach does not evaluate the effects of trt while accounting for across-subject heterogeneity.

3(c)

```
library(dplyr)
library(ggplot2)
subject_trt <- cd4 %>% select(pid,trt) %>% distinct()
subject_coeffs <- cd4 %>%
  group_by(pid) %>%
  do(model = lm(cd4 ~ time, data = .)) %>%
  summarise(intercept= coef(model)[1], slope = coef(model)[2], pid)
subject_coeffs <- left_join(subject_coeffs, subject_trt, by = "pid")
subject_coeffs_filtered <- subject_coeffs %>%
  filter(is.finite(intercept), is.finite(slope))
ggplot(subject_coeffs_filtered, aes(x = intercept, fill = as.factor(trt))) +
  geom_histogram(position = "identity", alpha = 0.5, bins = 30) +
  labs(title = "Histogram of Intercepts (beta_0,j)", x = "Intercept", y = "Frequency") +
  scale_fill_discrete(name = "Treatment") +
  theme_minimal()
```




```
ggplot(subject_coeffs_filtered, aes(x = slope, fill = as.factor(trt))) +
  geom_histogram(position = "identity", alpha = 0.5, bins = 30) +
  labs(title = "Histogram of Slopes (beta_1,j)", x = "Slope", y = "Frequency") +
  scale_fill_discrete(name = "Treatment") +
  theme_minimal()
```




```
t_test_intercept<- t.test(subject_coeffs$intercept ~ subject_coeffs$trt)
t_test_slope <- t.test(subject_coeffs$slope ~ subject_coeffs$trt)
print(t_test_intercept)
```

```
##
##  Welch Two Sample t-test
##
## data:  subject_coeffs$intercept by subject_coeffs$trt
## t = 0.64324, df = 247.27, p-value = 0.5207
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to
0
## 95 percent confidence interval:
##  -0.2510200  0.4944924
## sample estimates:
## mean in group 0 mean in group 1
##        4.826761        4.705025
```

```
print(t_test_slope)
```

```
##
##  Welch Two Sample t-test
##
## data:  subject_coeffs$slope by subject_coeffs$trt
## t = -0.52183, df = 203.03, p-value = 0.6024
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to
0
## 95 percent confidence interval:
##  -0.6799453  0.3953581
## sample estimates:
## mean in group 0 mean in group 1
##      -0.5530433      -0.4107497
```

From the two histograms, we see that both treatment groups have similar distributions. The t-test result gives us a p-value of 0.5207 for intercept and 0.6024 for slope, where both are larger than 0.05. This indicates that there is no significant difference between the intercepts and slops of the two treatment groups, which aligns with what we observe from the histograms.

There are several limitations of this model.

- There is a loss of power when we separate each subject. We have 254 subjects in total and we only have 1072 total data points. Then when we fit this model for each separate subject, there are approximately only 4-5 data points per subject. The models are very weak with this kind of sample size.
- The t-test does not account for possible random effects. We are essentially treating each subject's data in isolation. This means that any individual differences are not modeled in a systematic way. If some subjects naturally start with higher CD4 levels or respond differently to treatment, this variability should be included in the model. A mixed-effects model might be more appropriate.