Q2Q3

Hun Kang

2024-09-27

Q2.

```
dat = dget("https://www2.stat.duke.edu/~pdh10/Teaching/610/Homework/nels_math_ses")
dat$school = as.factor(dat$school)
str(dat)
```

```
## 'data.frame':    1993 obs. of  4 variables:
##  $ school   : Factor w/ 100 levels "1011","1031",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ mathdeg  : num  4 4 6 4 6 4 5 6 4 4 ...
##  $ mathscore: num  52.1 57.6 66.4 44.7 40.6 ...
##  $ ses      : num  -0.25 0.58 -0.85 -0.8 -1.41 -1.07 0.27 -0.16 -1 -1.22 ...
```

a.

```
mod = lm(mathscore ~ school, data = dat)
z = abs(mod$res)
anova(lm(z ~ dat$school))
```

```
## Analysis of Variance Table
##
## Response: z
##              Df Sum Sq Mean Sq F value  Pr(>F)
## dat$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
```

b.

```
mod = lm(mathscore ~ school*ses, data = dat)
z = abs(mod$res)
anova(lm(z ~ dat$school))
```

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

c.

Apparently `ses` explained a portion of variations in the residuals of each group, increasing p-value, favoring the null hypothesis of equal variance.

Q3.

```
dat = read.table(url("https://www2.stat.duke.edu/~pdh10/Teaching/610/Homework/cd4.dat"),header=TRUE)
head(dat)
```

```
##   pid  cd4 trt time
## 1   1 4.24   1 0.00
## 2   1 6.08   1 0.56
## 3   1 3.61   1 0.79
## 4   1 3.61   1 1.42
## 5   1 3.46   1 1.94
## 6   2 1.00   0 0.00
```

a.

```
fit0 = lm(cd4 ~ time, data = dat)
summary(fit0)
```

```
##
## Call:
## lm(formula = cd4 ~ time, data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.7517 -0.8093  0.2142  1.0605  4.6517
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.75171    0.07577  62.712   <2e-16 ***
## time        -0.20456    0.10013  -2.043   0.0413 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.563 on 1070 degrees of freedom
## Multiple R-squared:  0.003885,   Adjusted R-squared:  0.002954
## F-statistic: 4.174 on 1 and 1070 DF,  p-value: 0.0413
```

```r
fit1 <- lm(cd4 ~ time * trt, data = dat)
summary(fit1)
```

```
##
## Call:
## lm(formula = cd4 ~ time * trt, data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.7786 -0.7879  0.1943  1.0655  4.3375
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.85572    0.10988  44.190   <2e-16 ***
## time        -0.07557    0.14570  -0.519    0.604
## trt         -0.19257    0.15102  -1.275    0.203
## time:trt    -0.23227    0.19966  -1.163    0.245
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.555 on 1068 degrees of freedom
## Multiple R-squared:  0.01609,    Adjusted R-squared:  0.01332
## F-statistic:  5.82 on 3 and 1068 DF,  p-value: 0.0006053
```

```r
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
```

The model with `trt` has the fit statistically significantly better than the model without it. In this model, the slope of `time` is different according to `trt` level. However, we need to account for the potential across group heterogeneity.

b.

```r
fit2 = lm(cd4 ~ time* factor(pid), data = dat)
fit2b = lm(cd4 ~ time*factor(pid)+ time*trt, data=dat)
anova(fit2, fit2b)
```

```
## Analysis of Variance Table
##
## Model 1: cd4 ~ time * factor(pid)
## Model 2: cd4 ~ time * factor(pid) + time * trt
```

```
##   Res.Df    RSS Df Sum of Sq F Pr(>F)
## 1    598 300.79
## 2    598 300.79  0         0
```

`trt` and `pid` is confounded, that is, observations inside the same group all receive the same `trt` level. Therefore, the design matrix of two models span the column space, yielding the same model fits. For this reason, if we treat individual group effect as fixed effects, we cannot estimate `trt` as a fixed effect.

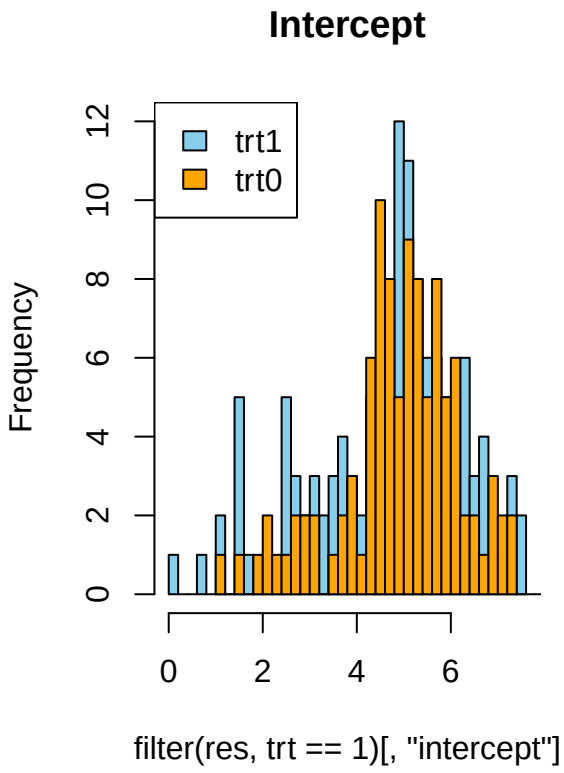
c.

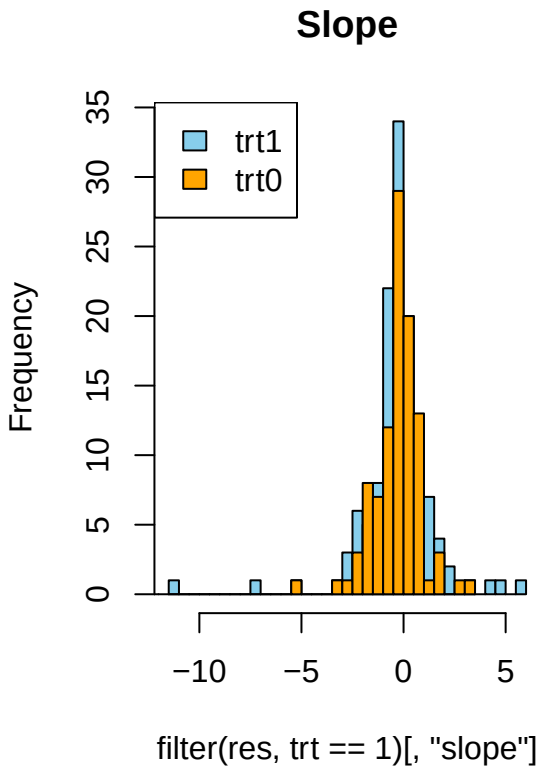
Two histograms do not differ a lot it appears, and t-test result is not significant. However, since each OLS estimates is from each group of very small sample size, we cannot expect this test to have enough statistical power.

```
library(tidyverse)
```

```
fits = dat %>%
  group_by(pid) %>%
  do(model = lm(cd4 ~ time, data=.))
J = length(fits$pid)
alphas = betas = trt = numeric(length=J)
for(j in 1:J){
  alphas[j] = coef(fits$model[[j]])[1]
  betas[j] = coef(fits$model[[j]])[2]
  trt[j] = filter(dat, pid == fits$pid[[j]])[1, "trt"]
}
res = na.omit(data.frame(intercept = alphas, slope = betas, trt = trt))
```

```
par(mfrow=c(1,2))
hist(filter(res, trt==1)[,"intercept"], breaks=30, main="Intercept", col="skyblue")
hist(filter(res, trt==0)[,"intercept"], breaks=30, main="Intercept", col="orange", add= T)
legend('topleft', c('trt1', 'trt0'), fill=c('skyblue', 'orange'))
hist(filter(res, trt==1)[,"slope"], breaks=30, main="Slope", col="skyblue")
hist(filter(res, trt==0)[,"slope"], breaks=30, main="Slope", col="orange", add= T)
legend('topleft', c('trt1', 'trt0'), fill=c('skyblue', 'orange'))
```






```
t.test(intercept ~ trt, data = res)
```

```
##
##  Welch Two Sample t-test
##
## data:  intercept by trt
## t = 1.4865, df = 221.8, p-value = 0.1386
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -0.09492523  0.67771077
## sample estimates:
## mean in group 0 mean in group 1
##        4.932246        4.640853
```

```
t.test(slope ~ trt, data = res)
```

```
##
##  Welch Two Sample t-test
##
## data:  slope by 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
```