Homework1

Jon Campbell

```
library(tidyverse)
library(knitr)
```

Part 1

Question 1

```
Y_obs <- c(8.62,1.48,8.93,9.57,2.65,7.3,.06,1.72,2.19,7.32,7.53,7.62)
Z <- c(rep(0,6), rep(1,6))
```

a)

```
Y_obs <- c(8.62,1.48,8.93,9.57,2.65,7.3,.06,1.72,2.19,7.32,7.53,7.62)
Z <- c(rep(0,6), rep(1,6))

tstat_obs <- mean(Y_obs[Z == 1]) - mean(Y_obs[Z == 0])

ind_combos <- combn(1:12,6)
tstats <- vector(mode = "double",length = ncol(ind_combos))

for (i in 1:ncol(ind_combos)) {
  Zperm <- rep(0,12)
  Zperm[ind_combos[,i]] = 1
  tstats[i] = mean(Y_obs[Zperm==1]) - mean(Y_obs[Zperm==0])
}

pval <- mean(abs(tstats) >= abs(tstat_obs))
```

The two-tailed p-value is 0.2706.

b)

```
set.seed(2929)
sampled_tstats <- sample(tstats, size = 1000, replace = TRUE)
pval_1000 <- mean(abs(sampled_tstats) >= abs(tstat_obs))
```

The two-tailed p-value from 1000 samples from the distribution under the Sharp Null Hypothesis is 0.27.

c)

```
ttest_pval <- t.test(Y_obs[Z==1], Y_obs[Z==0],var.equal = TRUE)$p.value
```

The p-value using a t-test is 0.3368.

d)

(b)'s approximation of (a) is part of the assignment mechanism component of the potential outcome framework as it draws from the distribution of all possible treatment assignments.

(c)'s approximation of (a) falls under the probabilistic model component of the potential outcome framework as it assumes the data in both groups is normally distributed with equal variance.

Question 2

a)

```
Y_obs_orig <- matrix(Y_obs,nrow = 6)
colnames(Y_obs_orig) <- c("0", "1")

combos <-expand.grid(pair1 = 0:1,pair2 = 0:1, pair3 = 0:1
                     ,pair4 = 0:1,pair5 = 0:1, pair6 = 0:1)
```

```
tstats <- vector(mode = "double",length = nrow(combos))
for (i in 1:nrow(combos)) {
  Y_obs_perm <- Y_obs_orig
  ind <- which(combos[i,] == 1)
  Y_obs_perm[ind,] = Y_obs_perm[ind,c(2,1)]
  tstats[i] = mean(Y_obs_perm[,2]) - mean(Y_obs_perm[,1])
}
```

```r
pval <- mean(abs(tstats) >= abs(tstat_obs))
```

The p-value from randomizing within pairs is 0.375.

b)

```r
set.seed(2121)
sampled_tstats <- sample(tstats, size = 1000, replace = TRUE)
pval_1000 <- mean(abs(sampled_tstats) >= abs(tstat_obs))
```

The p-value from sampling is 0.398.

c)

```r
ttest_pval <- t.test(Y_obs[Z==1], Y_obs[Z==0], var.equal = TRUE, paired = TRUE)
```

Using a paired t-test the p-value is 0.3652.

d)

Part (2b) is a part of the assignment mechanism as it makes sure $Z \perp X$ through randomization.

Part (2c) is a part of the probabilistic model just like question 1.

Question 3

$$Y_i^{obs} = Z_i Y_i(1) + (1 - Z_i) Y_i(0)$$

$$\hat{\tau} = \frac{1}{n_1} \sum_{i=1}^{n} Z_i Y_i^{obs} - \frac{1}{n_0} \sum_{i=1}^{n} (1 - Z_i) Y_i^{obs}$$

$$\hat{\tau} = \frac{1}{n_1} \sum_{i=1}^{n} Z_i Y_i(1) - \frac{1}{n_0} \sum_{i=1}^{n} (1 - Z_i) Y_i(0)$$

$$\text{Under CRE } Z_i \perp Y_i(0), Y_i(1) \text{ and } E[Y(z)] = Y(z)$$

$$E[\hat{\tau}] = \frac{1}{n_1} \sum_{i=1}^{n} E[Z_i] Y_i(1) - \frac{1}{n_0} \sum_{i=1}^{n} E(1 - Z_i) Y_i(0)$$

$$= \frac{1}{n_1} \sum_{i=1}^{n} \frac{n_1}{n} Y_i(1) - \frac{1}{n_0} \sum_{i=1}^{n} \frac{n_0}{n} Y_i(0)$$

$$= \frac{1}{n_1} \sum_{i=1}^{n} Y_i(1) - \frac{1}{n_0} \sum_{i=1}^{n} Y_i(0) = \tau$$

Question 4

Yes, it is possible for there to be evidence of a additive treatment effect under randomization. If this is the case we can change the Null Hypothesis from $Y_i(0) = Y_i(1)$ to $Y_i(0) - Y_i(1) = \tau$ and set τ equal to some fixed value. To impliment this with the data above we can do a two-sample t-test but set μ equal to some fixed value.

Part 2

Question 1

```
pot_outcomes <- matrix(c(35, 40, 45 ,55, 55 ,55, 65, 70, 25, 30, 45, 55, 60, 65, 75, 80, 3
colnames(pot_outcomes) <- c("Y1","Y0")

sample_ind = combn(1:12, 4)
assign_ind = combn(1:4, 2)
```

```
mean_diffs <- matrix(NA, nrow = ncol(sample_ind)
                     , ncol = ncol(assign_ind))

for (i in 1:ncol(sample_ind)) {
```

```
  indexes <- sample_ind[,i]
  sample <- pot_outcomes[indexes,]

  for (j in 1:ncol(assign_ind)) {

    assignment <- assign_ind[,j]
    assignment <- assign_ind[,sample(1:6,1)]
    Y0_obs <- sample[assignment,"Y0"]
    Y1_obs <- sample[-assignment,"Y1"]
    diff <- mean(Y1_obs) - mean(Y0_obs)
    mean_diffs[i,j] = diff
  }
}
```

From Simulation: 232.3602032

```
var1 <- sum((pot_outcomes[,"Y1"]-mean(pot_outcomes[,"Y1"]))^2)/11

var0 <- sum((pot_outcomes[,"Y0"]-mean(pot_outcomes[,"Y0"]))^2)/11

diff_means <- mean(pot_outcomes[,"Y1"]) - mean(pot_outcomes[,"Y0"])
var01 <- sum((pot_outcomes[,"Y1"] - pot_outcomes[,"Y0"] - diff_means)^2)/11
form <- var0/6 + var1/6 - var01/12
```

From Formula: 75.4261364

Question 2

```
variances <- matrix(NA, nrow = ncol(sample_ind)
                    , ncol = ncol(assign_ind))

for (i in 1:ncol(sample_ind)) {

  indexes <- sample_ind[,i]
  sample <- pot_outcomes[indexes,]

  for (j in 1:ncol(assign_ind)) {

    assignment <- assign_ind[,j]
    Y0_obs <- sample[assignment,"Y0"]
    Y1_obs <- sample[-assignment,"Y1"]
```

```
    sigma <- var(Y0_obs)/2 + var(Y1_obs)/2

    variances[i,j] = sigma
  }
}
```

Neyman Estimator of Variance: 230.2083333

Part 3

```
bestair <- readxl::read_xlsx("bestair640-1.xlsx", sheet = "data")
bestair <- bestair |>
  mutate(across(treatment_arm:sbp_6mo, ~ if_else(is.na(.x)
                                                 ,mean(.x,na.rm = TRUE),.x))) |>
  mutate(race = if_else(race == 2,1,0))
```

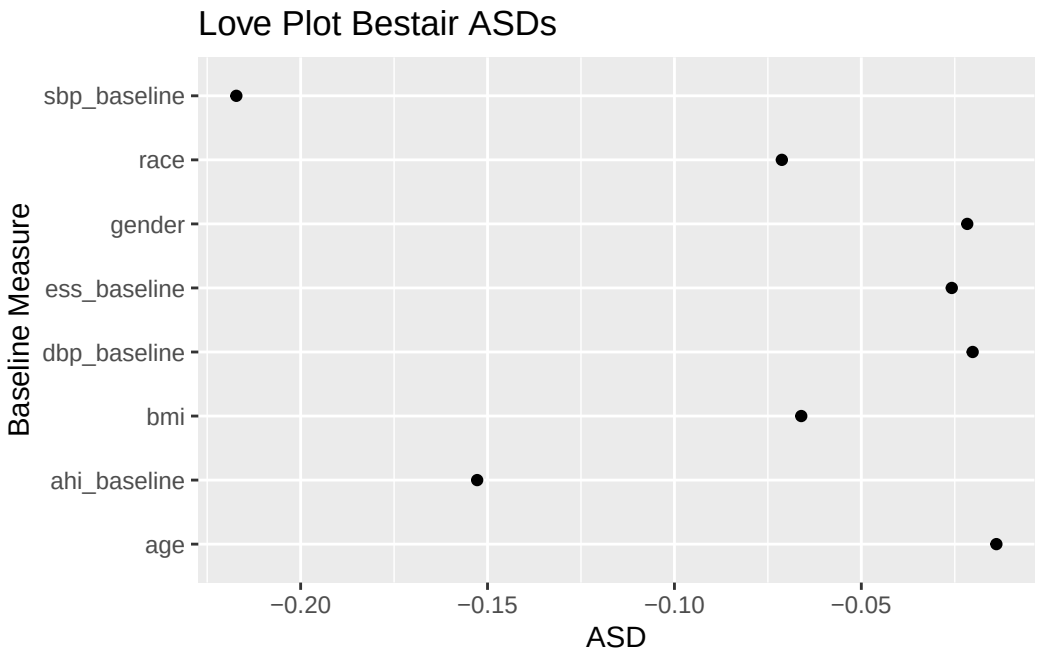
Question 1

```
baselines <- c("gender","age","bmi",
               "race","sbp_baseline","dbp_baseline","ahi_baseline","ess_baseline")
ASDs = matrix(NA, nrow = 1, ncol = 8)
colnames(ASDs) <- baselines
Z <- bestair$treatment_arm
for (bl in baselines) {
  X <- pull(bestair[,bl])
  s1 <- var(X[Z==1])
  s0 <- var(X[Z==0])
  diff_sum <- sum(X*Z)/sum(Z)-sum(X*(1-Z))/sum(1-Z)
  asd <- diff_sum/sqrt(s1+s0)
  ASDs[,bl] <- asd

}
```

```
#love plot
asd_dat <- tibble(
  bls = baselines,
  asd = ASDs[1,]
)
```

```
ggplot(asd_dat,aes(x = asd, y = bls)) +
  geom_point() +
  labs(title = "Love Plot Bestair ASDs"
       ,y = "Baseline Measure"
       ,x = "ASD")
```



Question 2

```
Z <- bestair$treatment_arm
Y <- bestair$sbp_6mo
estimate_unadj <- mean(Y[Z==1]) - mean(Y[Z==0])
se_unadj <- summary(lm(Y~Z))$coefficients[2,"Std. Error"]
hw_unadj <- sqrt(car::hccm(lm(Y ~ Z), type = "hc2")[2, 2])
unadj_res <- c(estimate_unadj,se_unadj,hw_unadj)
```

```
bestair_centered <- bestair|>
  mutate(across(gender:ess_baseline, ~ .x-mean(.x)))

ancova1 <- lm(formula = sbp_6mo~.,data = bestair_centered)
estimate_anc1 <- ancova1$coefficients[["treatment_arm"]]
```

```r
se_anc1 <- summary(ancova1)$coefficients[2,"Std. Error"]
hw_anc1 <-sqrt(car::hccm(ancova1
                         ,type = "hc2")["treatment_arm","treatment_arm"])
anc1_res <- c(estimate_anc1, se_anc1, hw_anc1)

ancova2 <- lm(formula = sbp_6mo~.^2,data = bestair_centered)
estimate_anc2 <- ancova2$coefficients[["treatment_arm"]]
se_anc2 <- summary(ancova2)$coefficients[2,"Std. Error"]
hw_anc2 <- NA
anc2_res <- c(estimate_anc2, se_anc2, hw_anc2)
```

```r
results <- cbind(unadj_res,anc1_res,anc2_res)
colnames(results) <- c("Unadjusted ATE","ANCOVA1","ANCOVA2")
rownames(results) <- c("Estimate","s.e.","Huber-White s.e.")
results |>
  round(3) |>
  kable()
```

	Unadjusted ATE	ANCOVA1	ANCOVA2
Estimate	-4.907	-2.540	-2.078
s.e.	2.230	1.683	2.092
Huber-White s.e.	2.266	1.696	NA

The Estimates decrease as we use a more flexible model and add interaction terms.

Question 3

```r
bestair_hyperten <- bestair_centered |>
  mutate(resist_hyperten = if_else(sbp_6mo>=130,1,0)) |>
  select(treatment_arm:ess_baseline,resist_hyperten)
```

a)

```r
Z <- bestair$treatment_arm
Y <- bestair_hyperten$resist_hyperten
bin_unadj_est <- mean(Y[Z==1]) - mean(Y[Z==0])
bin_ols <- lm(resist_hyperten~., data = bestair_hyperten)
bin_est <- bin_ols$coefficients["treatment_arm"]
```

```r
bin_ols_inter <- lm(resist_hyperten~.^2, data = bestair_hyperten)
bin_est_inter <- bin_ols_inter$coefficients["treatment_arm"]

results <- cbind(bin_unadj_est,bin_est,bin_est_inter)
colnames(results) <- c("Unadjusted ATE","ANCOVA1","ANCOVA2")
rownames(results) <- "Binary Estimates"
results |> round(3) |> kable()
```

	Unadjusted ATE	ANCOVA1	ANCOVA2
Binary Estimates	-0.208	-0.128	-0.107

b)

```r
logist_unadj <- glm(resist_hyperten~treatment_arm, family = binomial(link = "logit"), data

logist_anc1 <- glm(resist_hyperten~., family = binomial(link = "logit"), data = bestair_hy

logist_anc2 <- glm(resist_hyperten~.^2, family = binomial(link = "logit"), data = bestair_
```

```r
unadj_ate <- logist_unadj$coefficients[[2]]
anc1 <- logist_anc1$coefficients[["treatment_arm"]]
anc2 <- logist_anc2$coefficients[["treatment_arm"]]

results <- cbind(unadj_ate,anc1,anc2)
colnames(results) <- c("Unadjusted ATE","ANCOVA1","ANCOVA2")
rownames(results) <- "Logistic Estimates"
results |> round(3) |> kable()
```

	Unadjusted ATE	ANCOVA1	ANCOVA2
Logistic Estimates	-0.985	-0.818	-5.34

The linear model seems to perform better than logistic regression.