Exercise 2

This second example explores the use of statistical adjustment using a parametric model. For the purpose of the example, we always assume that the given covariates fulfill ignorability. We may have learned this through a DAG, and found the variables that fulfill the back-door criterion.

Here is a worked out, and commented example of using a parametric model and emmeans, for adjustment. For simplicity, the outcome variable is simply called Y, the treatment T, and the covariates W1-W5.

The data is simulated, and if you like you can ignore the simulation code. It is provided here for the sole reasons that you can reproduce the example quickly without having to download data files.

```
####simulation code##########################
set.seed(12345)
u <- rnorm(2000,0,1)
w1 <- 2*u + rnorm(2000,5,5)
w2 <- 2*u + rnorm(2000,5,5)
w3 <- 2*u + rnorm(2000,5,5)
p <- (1/(1+exp(-2+.2*w1 + .2*w2 - .3*w3 + .1*w1*w2)))
t <- rbinom(2000,1,p)
y <- 100 + 5*t + 2*w1 - 1*w2 - 1*w3 + .5*w1*w3 + .1*w1*t + rnorm(2000,0,2)
t <- factor(t)
levels(t) <- c("control","treatment")
############################################

####analysis code###########################

#unadjusted model
library(emmeans)
lm.u <- lm(y~t)
summary(lm.u)
```

```
##
## Call:
## lm(formula = y ~ t)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -91.458 -19.368  -5.216  12.025 187.843
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 122.9879     0.9202 133.658   <2e-16 ***
## ttreatment  -10.8560     1.3085  -8.296   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 29.26 on 1998 degrees of freedom
## Multiple R-squared:  0.0333, Adjusted R-squared:  0.03282
## F-statistic: 68.83 on 1 and 1998 DF,  p-value: < 2.2e-16
```

In a first step, I am looking at the unadjusted model. This is sometimes referred to as the prima facie effect (effect at first sight). I am using the summary statement in R. I do this for demonstration purposes, but I actually do not recommend it. Instead, I prefer using the emmeans statement, which estimates marginal means of the model, and compares them. We also see that the prima facie effect is biased.

```
summary(emmeans(lm.u,"t",contr="pairwise",weights="proportional"),infer=TRUE)
```

```
## $emmeans
##  t           emmean        SE   df lower.CL upper.CL t.ratio p.value
##  control   122.9879 0.9201703 1998 121.1833 124.7925 133.658  <.0001
##  treatment 112.1319 0.9303485 1998 110.3074 113.9565 120.527  <.0001
##
## Confidence level used: 0.95
##
## $contrasts
##  contrast            estimate       SE   df lower.CL upper.CL t.ratio
##  control - treatment 10.85597 1.308534 1998 8.289736 13.42221   8.296
##  p.value
##   <.0001
##
## Confidence level used: 0.95
```

Dissecting the emmeans command, we can see that the arguments are a) the lm model, b) the contrasts we want (here pairwise to compare group means), and c) a weights statement, which we typically set to proportional. The whole function is put into a summary statement with argument infer=TRUE. Note on the side: for folks who like the pipe operator in R, we could have also piped the lm statement into an emmeans statement, and that into a summary statement.

The emmeans output shows us first, the two group means (with inferentials), and then a group mean difference (here the estimate of the prima facie treatment effect).

We now repeat the same exercise, but we now adjust on all observed covariates, and re-estimate the model. I am again presenting the summary statement, and the emmeans statement, but I stronlgy suggest you to use emmeans for interpretation.

```
#linear adjustment on pre-test
lm.a <- lm(y~t+w1+w2+w3)
summary(lm.a)
```

```
##
## Call:
## lm(formula = y ~ t + w1 + w2 + w3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -66.797  -6.322  -1.491   5.060  94.537
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 89.30203    0.98313  90.835  < 2e-16 ***
## ttreatment   4.51209    0.95940   4.703 2.74e-06 ***
## w1           4.60206    0.07762  59.289  < 2e-16 ***
## w2          -1.05189    0.07713 -13.639  < 2e-16 ***
## w3           1.65169    0.06539  25.260  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.8 on 1995 degrees of freedom
## Multiple R-squared:  0.7529, Adjusted R-squared:  0.7524
## F-statistic:  1519 on 4 and 1995 DF,  p-value: < 2.2e-16
```

```r
summary(emmeans(lm.a,"t",contr="pairwise",weights="proportional"),infer=TRUE)
```

```
## $emmeans
##  t           emmean        SE   df lower.CL upper.CL t.ratio p.value
##  control   115.3884 0.5785005 1995 114.2539 116.5229 199.461  <.0001
##  treatment 119.9005 0.5871863 1995 118.7489 121.0520 204.195  <.0001
##
## Confidence level used: 0.95
##
## $contrasts
##  contrast            estimate        SE   df  lower.CL  upper.CL t.ratio
##  control - treatment -4.51209 0.9593966 1995 -6.393615 -2.630566  -4.703
##  p.value
##   <.0001
##
## Confidence level used: 0.95
```

Exercise:

1.) Download the file dfex2a from github (https://raw.githubusercontent.com/felixthoemmes/IPN_workshop/master/dfex2a.csv). You can download this file directly into R (no need to navigate to github in a browser, using the following code snippet:

```r
library(readr)
dfex2a <- read_csv("https://raw.githubusercontent.com/felixthoemmes/
                   IPN_workshop/master/dfex2a.csv")
```

The file contains a treatment t, an outcome y, and covariates x1-x5. We assume that these variables are those that fulfill the back-door criterion. Obtain an unadjusted estimate for the effect of t on y, using the emmean statement, and interpret the results.

2.) Now use a parametric model for adjustment, using covariates x1-x5. Obtain the adjusted treatment effect, and compare it to the unadjusted estimate.

3.) Now download the file dfex2b from github (https://raw.githubusercontent.com/felixthoemmes/IPN_workshop/master/dfex2b.csv). It contains a treatment t, and outcome y, and a single covariate m1. We again assume ignorability. This time, the treatment t has three levels (1,2,3). Those could e.g., be a control, and two active treatments. First obtain the unadjusted treatment effect among all pairwise groups. Use emmeans, and if you wish, you can contrast this with the regular summary command from the lm statement.

4.) Now use a parametric model again for adjustment of the effect. Use m1 as the single covariate, and obtain the adjusted pairwise effects. Again, you may wish to compare this to the summary statement.