Proctor Data Analysis Assessment

Chunyi Yu

2023-04-16

Background

From Won et al.(2017) Am J Trop Med Hyg: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5462587/

Schistosomiasis, caused by infection with Schistosoma spp., affects more than 200 million people worldwide. Prevalence and intensity of infection with Schistosoma mansoni peak between 10 and 15 years of age and gradually decline with age. In children, chronic schistosomiasis is associated with anemia and malnutrition and can compromise growth and cognitive development. Because of the influence school-aged children have on transmission of schistosomiasis, mass treatment of this age group with praziquantel has been the cornerstone of schistosomiasis control activities. Until recently, disease burden and morbidity among preschool-aged children have remained understudied. However, recent research has shown that first infection is often acquired at a very young age, and there is growing evidence that the burden of disease among PSAC may warrant global attention.

A field study was conducted from 2012 to 2014 in Mbita subcounty, which borders Lake Victoria in western Kenya. Before the start of the study, malaria interventions had been in place for several years, but no mass drug administration (MDA) for schistosomiasis had been conducted.

Thirty villages that met the selection criteria were randomized into two study arms to compare different MDA strategies for schistosomiasis and STH programs. Fifteen villages were randomized to a community-wide treatment arm and the remaining 15 villages were randomized to a school-based treatment arm. In each of the 30 study villages, the study aimed to enroll 100 preschool aged children (1~5 years) and their mothers or guardians.

In both study arms, parasitologic and serologic indicators were monitored at baseline (year 1 in 2012) and annually following treatment. All annual monitoring was done using repeated, cross-sectional surveys in the selected villages and children were treated with praziquantal and albendazole approximately 2 months after each annual measurement.

1 Download and process the data (data cleaning)

1. Review the data and codebooks. Familiarize yourself with them. Then, read in the data. (These data are very clean so unlike most projects there is not an extensive amount of data processing required. In reality, data processing is often the most time consuming part of a project!) Depending on how you approach the work below, if you want to join the two datasets the 1:many key variable is village id (vid).
2. Create a derived variable that is an indicator of whether the child was seropositive to either the SEA antigen or the Sm25 antigen. We will be using a combined measure of serpositivity as the outcome!

```
  kids  <- read.table ("C:/Users/fjyuc/Desktop/DATAGAME/R/Assessment_proctor_UCSF_Grace/test/mbita_schisto.csv", header=TRUE, sep=",")
  villages  <- read.table ("C:/Users/fjyuc/Desktop/DATAGAME/R/Assessment_proctor_UCSF_Grace/test/mbita_spatial.csv", header=TRUE, sep=",")

  library(dplyr)
  complete<-left_join(kids,villages, by="vid")
  complete$serop_indi<-ifelse(complete$sea_pos ==1|complete$sm25_pos==1,1,0)
```

```
  #transform to factor
  complete <- transform(complete,
                        vid=factor(vid),
                        year=factor(year,levels=c(2012, 2013, 2014), labels=c("2012","2013"
,"2014")),
                        arm=factor(arm,levels=c("CWT","SBT"),labels=c("CWT","SBT")),
                        sex=factor(sex,levels=c("male","female"), labels=c("M","F")),
                        sea_pos=factor(sea_pos,levels=c(0,1),labels=c("No","Yes")),   #  se
ropositive to the SEA antigen;
                        sm25_pos=factor(sm25_pos,levels=c(0,1),labels=c("No","Yes")), # ser
opositive to the Sm25 antigen;
                        kk_pos=factor(kk_pos,levels=c(0,1),labels=c("No","Yes")), # Kato-Ka
tz positive for S. mansoni;
                        serop_indi=factor(serop_indi,levels=c(0,1),labels=c("No","Yes"))) #
seropositive to SEA or Sm25.

    label(complete$vid)       <- "Village ID"
    label(complete$pid)       <- "Child ID"
    label(complete$year)      <- "Study year"
    label(complete$arm)       <- "Study arm"
    label(complete$agey)      <- "Age"
    label(complete$sex)       <- "Sex"
    label(complete$sea)       <- "Sea response"
    label(complete$sea_pos)   <- "Sea Positive"
    label(complete$sm25)      <- "Sm25 response"
    label(complete$sm25_pos)  <- "Sm25 Positive"
    label(complete$sm_epg)    <- "Eggs per gram"
    label(complete$kk_pos)    <- "KK Positive"
    label(complete$elev)      <- "Village elevation"
    label(complete$tmin)      <- "Average minimum temperature"
    label(complete$prec)      <- "Average precipitation"
    label(complete$dist_victoria)   <- "Distance to Lake Victoria"
    label(complete$serop_indi)      <- "A seropositivity indicator"

    units(complete$agey) <- "years"
    units(complete$elev) <- "meters"
    units(complete$tmin) <- "F"
    units(complete$prec) <- "mm"
    units(complete$dist_victoria) <- "meters"

    attach(complete)
```

This is a data containing information from a study on Schistosoma mansoni infection in children from 30 villages in Kenya over three years of 2012~2014. The dataset contains 3663 observations and 17 variables. Each observation represents a child who participated in the study. The variables include the study year, village ID, study arm, individual child ID, age, sex, measures of infection (SEA response, Sm25 response, and Kato-Katz eggs per gram of stool), seropositivity to SEA and Sm25 antigens, and Kato-Katz results. It also includes information on village elevation, average minimum temperature, average precipitation, distance to Lake Victoria. Finally, there is an indicator variable for whether the child was seropositive to either the SEA or Sm25 antigens, which is the primory outcome for this study.

#2 Describe the data (numeric summaries with data visualization)

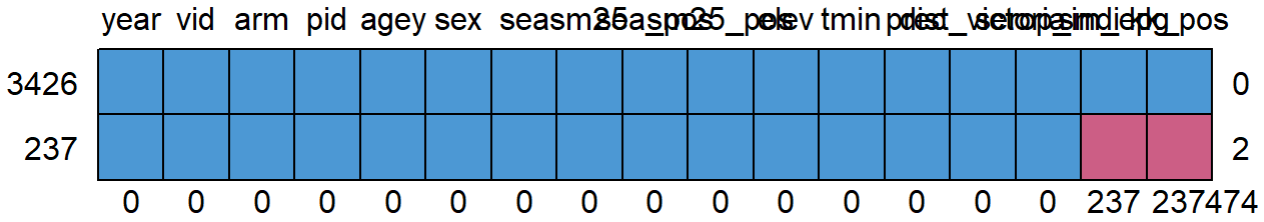
Provide some simple descriptive summaries that help describe the data.
For example, how many children were measured at baseline (2012) and in subsequent survey visits based on blood-

and stool-based measures of infection? How complete were the measurements? For quantitative variables, is there anything you see about their distribution that might be important to consider in any downstream analysis? There is no one way to do this and we are not looking for a specific result. We are interested to learn how you would describe the data to have a sense for what it contains.

```
# 1. how many children were measured at baseline (2012) and in subsequent survey visits based
on blood- and stool-based measures of infection?

    library(dplyr)
    library(plyr)
    ddply(complete, c("sea","sm25","sm_epg")~ year,nrow)
```

```
##   c("sea", "sm25", "sm_epg") year  V1
## 1                       sea 2012 374
## 2                      <NA> 2013 395
## 3                      <NA> 2014 452
## 4                    sm_epg 2012 373
## 5                      <NA> 2013 396
## 6                      <NA> 2014 452
## 7                      sm25 2012 373
## 8                      <NA> 2013 396
## 9                      <NA> 2014 452
```

```
    #The first table shows the number of children measured at baseline and in the following t
wo years for three different measures of infection. For the blood-based measure of infection
with SEA, there were 374 children measured at baseline, 395 in the second year, and 452 in t
he third year. Similarly, for the blood-based measure of infection with sm_25, there were 37
3 children measured at baseline, 396 in the second year, and 452 in the third year. Lastly, f
or the stool-based measure of infection with egg per gram, there were 373 children measured
at baseline, 396 in the second year, and 452 in the third year.



# 2.How complete were the measurements?
  library(mice)
  md.pattern(complete)
```




```
##      year vid arm pid agey sex sea sm25 sea_pos sm25_pos elev tmin prec
## 3426    1   1   1   1    1   1   1    1       1        1    1    1    1
## 237     1   1   1   1    1   1   1    1       1        1    1    1    1
##         0   0   0   0    0   0   0    0       0        0    0    0    0
##      dist_victoria serop_indi sm_epg kk_pos
## 3426             1          1      1      1   0
## 237              1          1      0      0   2
##                  0          0    237    237 474
```

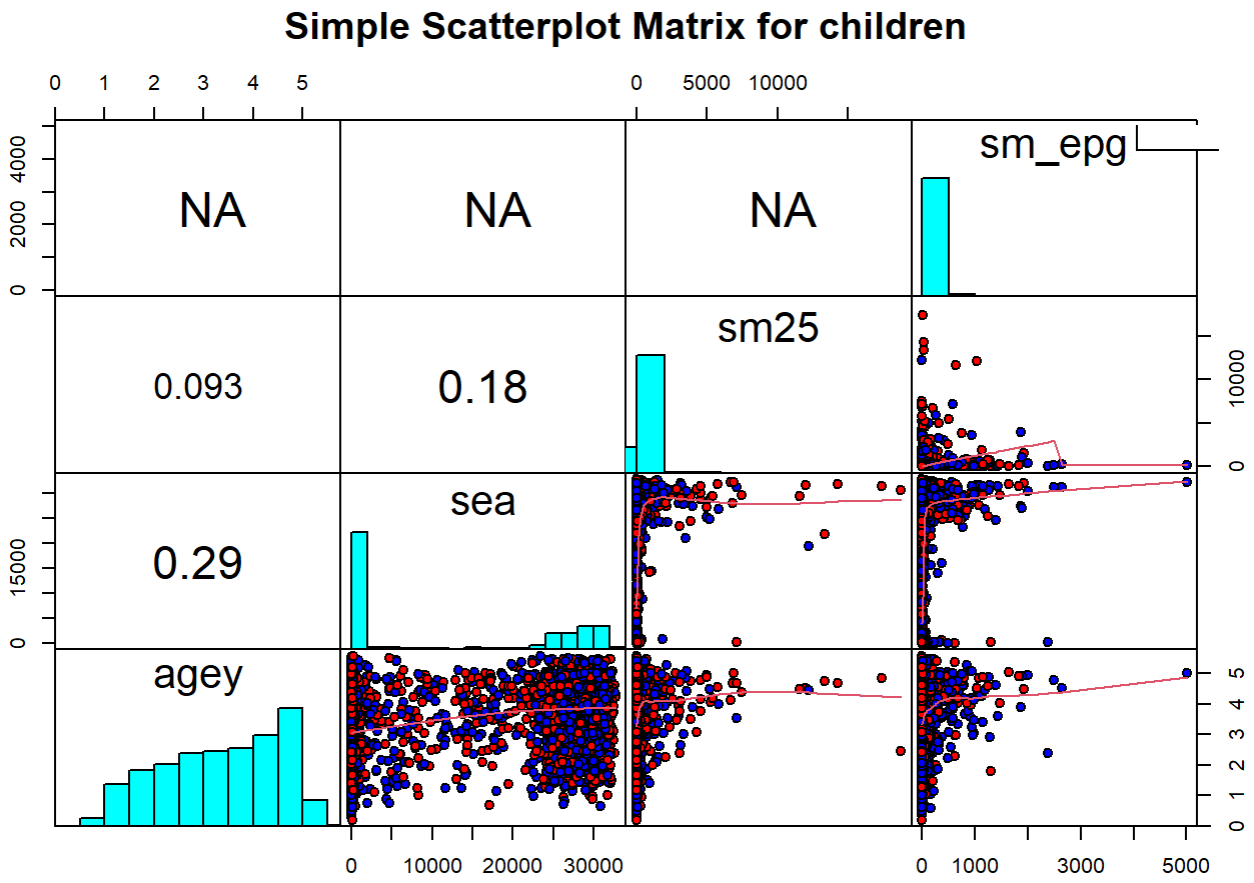
```
  summary(complete)
```

```
##    year           vid        arm            pid            agey     sex
##  2012:1120   18     : 196   CWT:1826   Min.   :   1   Min.   :0.2   M:1759
##  2013:1187   3      : 192   SBT:1837   1st Qu.: 916   1st Qu.:2.4   F:1904
##  2014:1356   8      : 187              Median :1832   Median :3.5
##              11     : 179              Mean   :1832   Mean   :3.4
##              12     : 169              3rd Qu.:2748   3rd Qu.:4.4
##              7      : 168              Max.   :3663   Max.   :5.6
##              (Other):2572
##       sea              sm25           sm_epg     sea_pos    sm25_pos
##  Min.   :    4   Min.   :  -12   Min.   :   0   No :1914   No :3040
##  1st Qu.:  101   1st Qu.:    1   1st Qu.:   0   Yes:1749   Yes: 623
##  Median :  387   Median :    6   Median :   0
##  Mean   :11982   Mean   :  129   Mean   :  42
##  3rd Qu.:27337   3rd Qu.:   19   3rd Qu.:  12
```

```
##  Max.   :32685   Max.   :18816   Max.   :5004
##                                  NA's   :237
##   kk_pos          elev           tmin          prec     dist_victoria
##  No  :2557   Min.   :1137   Min.   :149   Min.   : 81   Min.   :  32
##  Yes : 869   1st Qu.:1151   1st Qu.:157   1st Qu.: 83   1st Qu.: 234
##  NA's: 237   Median :1158   Median :159   Median : 85   Median : 713
##              Mean   :1170   Mean   :159   Mean   : 88   Mean   :1011
##              3rd Qu.:1178   3rd Qu.:161   3rd Qu.: 94   3rd Qu.:1598
##              Max.   :1338   Max.   :161   Max.   :101   Max.   :4718
##
##  serop_indi
##  No :1799
##  Yes:1864
##
##
##
##
##
```

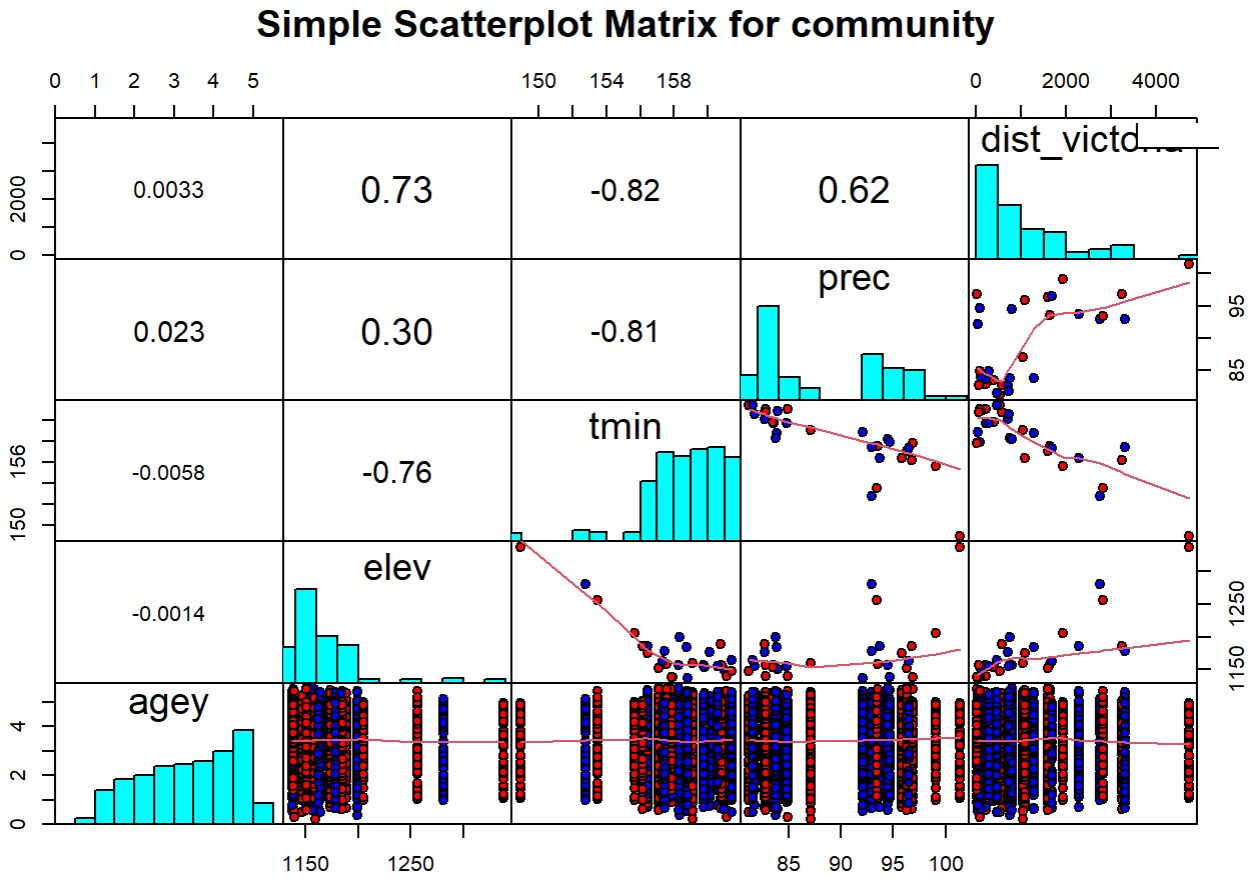
```
  # Based on the data visualization, it is apparent that there are 237 instances where values are missing in the variables of sm_epg and kk_pos simultaneously. This may be due to these 237 individuals being unable to take the sm_epg tests, resulting in the missing values.




#3. For quantitative variables, is there anything you see about their distribution that might be important to consider in any downstream analysis?

    # Create spaghetti plot
    panel.hist <- function(x, ...)
    {
        usr <- par("usr")
        par(usr = c(usr[1:2], 0, 1.5) )
        h <- hist(x, plot = FALSE)
        breaks <- h$breaks; nB <- length(breaks)
        y <- h$counts; y <- y/max(y)
        rect(breaks[-nB], 0, breaks[-1], y, col = "cyan", ...)
    } # put histograms on the diagonal


    panel.cor <- function(x, y, digits = 2, prefix = "", cex.cor, ...)
    {
        par(usr = c(0, 1, 0, 1))
        r <- cor(x, y)
        txt <- format(c(r, 0.123456789), digits = digits)[1]
        txt <- paste0(prefix, txt)
        if(missing(cex.cor)) cex.cor <- 0.8/strwidth(txt)
        text(0.5, 0.5, txt, cex = cex.cor * 0.4)
    }# put correlations on the upper panels

    pairs(~agey+sea+sm25+sm_epg,
          main="Simple Scatterplot Matrix for children",
```

```
        pch = 21,
        # panel = panel.smooth,
        lower.panel = panel.smooth,
        upper.panel = panel.cor,
        gap=0,
        row1attop=FALSE,
        diag.panel = panel.hist,
        bg = c("red", "blue")[unclass(complete$arm)])

  legend("topright", c("CWT", "SBT"), col= c("red", "blue"), pch=1)
```



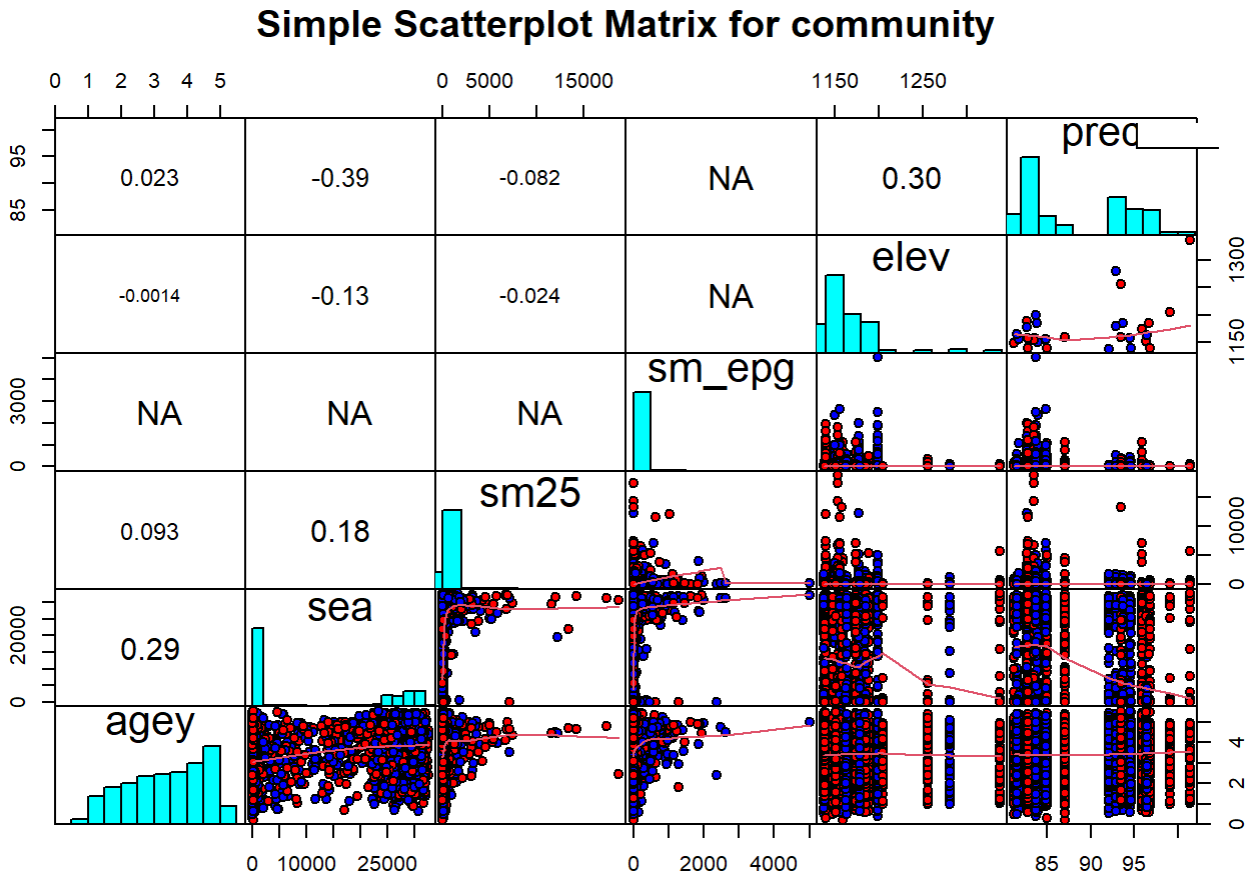

```
    # The distribution of the four variables can be summarized as follows: the age variable i
s not strongly skewed, but there are fewer observations in the youngest and oldest groups; th
e sea variable shows some extreme values on both ends, while the sm25 and sm_epg variables ar
e skewed to the right. Further investigation is required for observations with extreme values
in sea, sm25, and sm_epg. There is no strong correlation between the four variables, but ther
e is a weak correlation between age and sea, indicating that younger people tend to have high
er levels of sea. As for the relationship between age and the other three variables, higher r
esponse rates are mainly observed in younger age groups. The distribution between the treatme
nt and control groups (blue vs red) on the graph appears to be relatively even, but further s
tatistical comparisons are necessary for confirmation.


    pairs(~agey+elev+tmin+prec+dist_victoria,
          main="Simple Scatterplot Matrix for community",
          pch = 21,
```

```
        # panel = panel.smooth,
        lower.panel = panel.smooth,
        upper.panel = panel.cor,
        gap=0,
        row1attop=FALSE,
        diag.panel = panel.hist,
        bg = c("red", "blue")[unclass(complete$arm)])

  legend("topright", c("CWT", "SBT"), col= c("red", "blue"), pch=1)
```




```
# The four geographical variables show a strong correlation, particularly the distance from t
he lake (dist_victoria) which is strongly correlated with the other three variables with all
correlation coefficients exceeding 0.5. The variable is negatively correlated with the minimu
m temperature, meaning that the farther the distance from the lake, the lower the temperature
. There is also a strong negative correlation between temperature and the other two variables
of altitude and precipitation: the higher the altitude, the lower the temperature, and the mo
re precipitation. However, there is no clear relationship between altitude and precipitation.
The strong correlation among variables may lead to collinearity problems when building models
, which should be addressed.On the other hand, there is a weak correlation between age and th
e four geographical variables, suggesting that children of all ages are distributed across ea
ch region.


    pairs(~agey+sea+sm25+sm_epg+elev+prec,
        main="Simple Scatterplot Matrix for community",
        pch = 21,
```

```
        # panel = panel.smooth,
        lower.panel = panel.smooth,
        upper.panel = panel.cor,
        gap=0,
        row1attop=FALSE,
        diag.panel = panel.hist,
        bg = c("red", "blue")[unclass(complete$arm)])

  legend("topright", c("CWT", "SBT"), col= c("red", "blue"), pch=1)
```




```
# After reviewing the previous plots, two environmental variables and two variables related t
o the children were selected to create a graph. The graph shows a weak negative correlation b
etween altitude and the child-related variables. This suggests that the higher the altitude,
the lower the positive response, although the correlation is so weak that it can be ignored.
Additionally, there is a weak negative correlation between precipitation and the sea variable
. This indicates that the more precipitation there is, the lower the positive response value
of the sea variable. The relationship between these variables and the other variables is not
strong.



#4. Provide some simple descriptive summaries that help describe the data.

    # Group comparison by arm
    library(CBCgrps)
```

```
twogrps(complete[,-c(2,4)], "arm")
```

```
## $Table
##
## 1                      Variables         Total (n = 3663)
## 2                    year, n (%)
## 3                           2012                1120 (31)
## 4                           2013                1187 (32)
## 5                           2014                1356 (37)
## 6           agey, Median (Q1,Q3)         3.5 (2.43, 4.44)
## 7                     sex, n (%)
## 8                             M                1759 (48)
## 9                             F                1904 (52)
## 10           sea, Median (Q1,Q3)         387 (101, 27337)
## 11          sm25, Median (Q1,Q3)             6 (1.15, 19)
## 12        sm_epg, Median (Q1,Q3)                0 (0, 12)
## 13                sea_pos, n (%)
## 14                           No                1914 (52)
## 15                          Yes                1749 (48)
## 16              sm25_pos, n (%)
## 17                           No                3040 (83)
## 18                          Yes                 623 (17)
## 19                kk_pos, n (%)
## 20                           No                2557 (75)
## 21                          Yes                 869 (25)
## 22          elev, Median (Q1,Q3)        1158 (1151, 1178)
## 23          tmin, Median (Q1,Q3)  158.83 (157.42, 160.75)
## 24          prec, Median (Q1,Q3)      84.92 (83.42, 94.5)
## 25 dist_victoria, Median (Q1,Q3) 712.82 (234.21, 1598.12)
## 26             serop_indi, n (%)
## 27                           No                1799 (49)
## 28                          Yes                1864 (51)
##
## 1            CWT (n = 1826)           SBT (n = 1837)       p
## 2                                                      0.003
## 3                  538 (29)                 582 (32)
## 4                  563 (31)                 624 (34)
## 5                  725 (40)                 631 (34)
## 6          3.5 (2.45, 4.44)          3.5 (2.4, 4.45)    0.87
## 7                                                      0.367
## 8                  891 (49)                 868 (47)
## 9                  935 (51)                 969 (53)
## 10        227.5 (94, 26362)        4326 (113, 27899) < 0.001
## 11                5 (1, 18)                6 (2, 19)   0.191
## 12                 0 (0, 0)                0 (0, 12) < 0.001
## 13                                                     < 0.001
## 14                1042 (57)                 872 (47)
## 15                 784 (43)                 965 (53)
## 16                                                     0.534
## 17                1523 (83)                1517 (83)
## 18                 303 (17)                 320 (17)
## 19                                                     < 0.001
## 20                1315 (77)                1242 (72)
```

```
## 21                 386 (23)                 483 (28)
## 22        1156 (1147, 1174)        1164 (1155, 1184) < 0.001
## 23  159.08 (156.42, 160.83)  158.83 (158.25, 160.08)   0.021
## 24     87.08 (82.83, 95.83)     83.92 (83.58, 92.92) < 0.001
## 25 1046.1 (234.21, 1598.12) 712.82 (281.22, 1291.54)   0.008
## 26                                                     < 0.001
## 27                 989 (54)                 810 (44)
## 28                 837 (46)                1027 (56)
##
## $VarExtract
##  [1] "year"          "sea"           "sm_epg"        "sea_pos"
##  [5] "kk_pos"        "elev"          "tmin"          "prec"
##  [9] "dist_victoria" "serop_indi"
```

```
    #Another method for group comparisions with test types & P values
    library(epiDisplay)
    library(data.table)
    dt<-as.data.table(complete)
    tableStack(vars=c(year,agey,sea, sm25,sm_epg,sea_pos,sm25_pos,kk_pos,elev,tmin,prec,dist_v
ictoria), by = arm, dataFrame = dt)
```

```
##              CWT                  SBT                  Test stat.
## Total        1826                 1837
##
## year                                                   Chisq. (2 df) = 11.35
##    2012      538 (29.5)           582 (31.7)
##    2013      563 (30.8)           624 (34)
##    2014      725 (39.7)           631 (34.3)
##
## agey                                                   Ranksum test
##   median(IQR) 3.5 (2.4,4.4)       3.5 (2.4,4.5)
##
## sea                                                    Ranksum test
##   median(IQR) 227.5 (94,26362)    4326 (113,27899)
##
## sm25                                                   Ranksum test
##   median(IQR) 5 (1,18)            6 (2,19)
##
## sm_epg                                                 Ranksum test
##   median(IQR) 0 (0,0)             0 (0,12)
##
## sea_pos                                                Chisq. (1 df) = 33.8
##    No         1042 (57.1)          872 (47.5)
##    Yes        784 (42.9)           965 (52.5)
##
## sm25_pos                                               Chisq. (1 df) = 0.44
##    No         1523 (83.4)          1517 (82.6)
##    Yes        303 (16.6)           320 (17.4)
##
## kk_pos                                                 Chisq. (1 df) = 12.74
##    No         1315 (77.3)          1242 (72)
##    Yes        386 (22.7)           483 (28)
##
```

```
## elev                                                    Ranksum test
##   median(IQR) 1156 (1147,1174)      1164 (1155,1184)
##
## tmin                                                    Ranksum test
##   median(IQR) 159.1 (156.4,160.8)   158.8 (158.2,160.1)
##
## prec                                                    Ranksum test
##   median(IQR) 87.1 (82.8,95.8)      83.9 (83.6,92.9)
##
## dist_victoria                                           Ranksum test
##   median(IQR) 1046.1 (234.2,1598.1) 712.8 (281.2,1291.5)
##
##               P value
## Total
##
## year          0.003
##    2012
##    2013
##    2014
##
## agey          0.87
##   median(IQR)
##
## sea           < 0.001
##   median(IQR)
##
## sm25          0.191
##   median(IQR)
##
## sm_epg        < 0.001
##   median(IQR)
##
## sea_pos       < 0.001
##    No
##    Yes
##
## sm25_pos      0.506
##    No
##    Yes
##
## kk_pos        < 0.001
##    No
##    Yes
##
## elev          < 0.001
##   median(IQR)
##
## tmin          0.021
##   median(IQR)
##
## prec          < 0.001
##   median(IQR)
##
## dist victoria 0.008
```

```
##   median(IQR)
##
```

```
# A total of 3663 preschool-aged children were enrolled in the study during 2012 to 2014. Of
those enrolled, there were 1826 (49.84%) children in the CWT group, median age of enrollment
was 3.5 years (2.43, 4.44, P=0.87) for the total group: for CWT group, the median age was 3.5
(2.45, 4.44), and for the SBT group, it was 3.5 (2.4, 4.45); There were slightly more female
enrolled in the group (52%), with 51% for CWT and 53% for SBT. The overall prevalence of S.M
ansoni infection with antibody responses to SEA was lower in CWT (43%) compared to SBT (53%)
with P< 0.001; the overall prevalence of S.Mansoni infection by Kato-Katz was lower in CWT (2
3%) compared to SBT (28%) with P< 0.001; In contrast to the SEA results, prevalence of S.Mans
oni infection with antibodies to Sm25 between two groups - CWT (17%) and SBT (17%)- are very
similar.
```

#3 Summarize baseline characteristics (group comparisons:table 1)

In randomized, controlled trials (RCTs) we use random allocation of treatment to balance measurable and unmeasurable characteristics between treatment groups. On average, the potential outcomes in the two groups should be the same in the absence of treatment any differences we observe in outcomes can be attributed to a treatment effect.One important step in an RCT is to compare groups based on measurable baseline characteristics. This is often Table 1 in reporting for RCTs and is item 15 on the CONSORT checklist for cluster randomized trials.The schistosomiasis study was a community randomized trial so the independent units for analysis are the community.

1. Create a table that summarizes individual-level and cluster-level characteristics by randomized group (community-wide treatment and school-based treatment). Each row should be a variable or level of that variable. There should be a separate column for each group. For measures of S. mansoni infection or antibody response, limit your summary to the categorical measures rather than quantitative measures. For other quantitative variables, summarize the mean and standard deviation (and/or median and interquartile range if you feel that is more appropriate). For categorical variables, report the N and percent.

```
#  Individual level VS community level-----------------------------------------------------
-----------

# a table showing baseline characteristics for cluster and individual participant levels as a
pplicable for each group

    # Create a new data set for tables.
    data4comp <- complete[, -c(4)]

    library(dplyr)
    aggdata <- aggregate(data4comp[, c(4,6,7,8)], by = list(data4comp$vid), FUN = mean , na.r
m = T)
    colnames(aggdata)[1:5] <- c("vid","agem","seam","sm25m","sm_epgm")
    data4comp<-left_join(data4comp,aggdata, by="vid")
    data4comp<-data4comp[,-2]

    label(data4comp$agem)  <- "community Average age"
    label(data4comp$seam)  <- "community Sea response"
    label(data4comp$sm25m)  <- "community Sm25 response"
    label(data4comp$sm_epgm)  <- "community Eggs per gram"
    units(data4comp$agem)  <- "years"

    #Use table 1 to create the baseline tables
```

```
    #Functions and arguments
    my.render.cont <- function(x) {
        with(stats.apply.rounding(stats.default(x), digits=2), c("",
            "Mean (SD)"=sprintf("%s (&plusmn; %s)", MEAN, SD)))
     }
    my.render.cat <- function(x) {
        c("", sapply(stats.default(x), function(y) with(y,
            sprintf("%d (%0.0f %%)", FREQ, PCT))))
    }


    caption1  <- "Individual-level Baseline Characteristics by Randomized Group"
    footnote <- " "
    table1(~ year + agey + sex +sea_pos+sm25_pos+kk_pos| arm,
            overall=F,
            data=data4comp,
            caption=caption1,
            footnote=footnote,
            render.continuous=my.render.cont,
            render.categorical=my.render.cat)
```

Individual-level Baseline Characteristics by Randomized Group

	CWT (N=1826)	SBT (N=1837)
Study year		
2012	538 (29 %)	582 (32 %)
2013	563 (31 %)	624 (34 %)
2014	725 (40 %)	631 (34 %)
Age (years)		
Mean (SD)	3.4 (± 1.2)	3.4 (± 1.2)
Sex		
M	891 (49 %)	868 (47 %)
F	935 (51 %)	969 (53 %)
Sea Positive		
No	1042 (57 %)	872 (47 %)
Yes	784 (43 %)	965 (53 %)
Sm25 Positive		
No	1523 (83 %)	1517 (83 %)
Yes	303 (17 %)	320 (17 %)
KK Positive		
No	1315 (72 %)	1242 (68 %)
Yes	386 (21 %)	483 (26 %)
Missing	125 (6.8%)	112 (6.1%)

```
    caption2  <- "Cluster-level Baseline Characteristics by Randomized Group"
```

```
    footnote <- " "
    table1(~ agem +elev+tmin+prec+dist_victoria | arm,
            overall=F,
            data=data4comp,
            caption=caption2,
            footnote=footnote,
            render.continuous=my.render.cont,
            render.categorical=my.render.cat)
```

Cluster-level Baseline Characteristics by Randomized Group

	CWT (N=1826)	SBT (N=1837)
community Average age (years)		
Mean (SD)	3.4 (± 0.13)	3.4 (± 0.14)
Village elevation (meters)		
Mean (SD)	1200 (± 40)	1200 (± 29)
Average minimum temperature (F)		
Mean (SD)	160 (± 2.8)	160 (± 1.9)
Average precipitation (mm)		
Mean (SD)	89 (± 6.5)	88 (± 5.3)
Distance to Lake Victoria (meters)		
Mean (SD)	1100 (± 1100)	940 (± 920)

```
    # caption3  <- "Comparison of Baseline Characteristics on Cluster-level and Individual Le
vel"
    #  table1(~ agey + agem +sea+sm25 +sm_epg +seam+sm25m +sm_epgm | arm,
    #         overall=F,
    #         data=data4comp,
    #         caption=caption2,
    #         footnote=footnote,
    #         render.continuous=my.render.cont,
    #         render.categorical=my.render.cat)
```

2. Do the groups look well balanced at baseline based on measured characteristics? Briefly explain why you think they are or are not well balanced. The groups are basically well balanced at baseline based on measured characteristics.The total number of clusters for the treatment and control group are equal (both are 15); Based on the characteristics shown in both tables, and the break-down percentages of each variable/level between two groups appears to be quite similar between the two groups.

Note that the mean(sd) of age on the individual level is different from that on the community level, though the means are the same, sd on the individual level is much larger than on the community level. The difference in the variance should be taken account into model building.

3. Idea for reflection: what element of the design could contribute to better or worse balance between groups in their baseline characteristics? One element of the design could have great impact on the balance between groups is the number of clusters in each group. Different from a completely randomized control design, when we randomize groups of people instead of individuals, there is greater risk that the groups will end up being different from each other by chance, even if we did the randomization correctly. This is because there are usually fewer groups than individuals, especially the number of clusters is often small, so the randomization might not work out perfectly.

#4 Compare S. mansoni seroprevalence between groups (statistical modeling)

1.Estimate the effect of Community-Wide Treatment (CWT) versus School-Based Treatment (SBT) on S. mansoni seroprevalence as measured by IgG seropositivity to the Soluble Egg Antigen (SEA) and/or the recombinant Sm25 antigen. Since SBT is the current standard of care, treat that as the comparison group and CWT as the intervention group.

Compare groups based on an absolute measure of effect, namely the difference in prevalence, averaged over the entire post-treatment period (combining measurements over 2013 and 2014). Provide estimates of effect, 95% confidence intervals, and a P-value for the difference. Summarize your results in a formatted table. Provide a brief interpretation of the results.

There are multiple correct ways to do the analysis, but whatever approach you use remember: the independent unit in the trial is the community. We recommend either an analysis based on community-level means or an analysis based on child-level outcomes that accounts for outcome correlation within the community.

```
    # An analysis based on community-level means

    ######################
    # create a new dataset for modeling
    ######################

    # create an indicator variable serop_indi;
    d2 <- kids %>% mutate(serop_indi<-ifelse(sea_pos ==1|sm25_pos==1,1,0))

    # Transform the format of variables
    d3 <- d2 %>%
      ungroup() %>%
      mutate(vid=factor(vid),
             yearf = factor(year),
             serop =as.numeric(serop_indi)-1)


    # Create the response variable
    # calculate prevalence by village over 2013 and 2014 years
    detach(package:plyr)

    dt4model <- subset(d3, year!=2012) %>%
      group_by(vid) %>%
      summarize(serop_n = sum(serop,na.rm=T),
                serop_N = sum(ifelse(!is.na(serop),1,0))
                ) %>%
      mutate(serop_prev = serop_n/serop_N)

    # Create the final data set
    dt4model<-left_join(complete, dt4model, by="vid")
    dt4model<-subset(dt4model, select = c(year, vid,arm,agey,sex,elev,tmin,prec,dist_victoria,
serop_prev))

    #scale the large value variables
     dt4model<- transform(dt4model,
               selev=scale(elev, center = F,scale = T),
               stmin = scale(tmin, center = F,scale = T),
               sprec =scale(prec, center = F,scale = T),
```

```
                sdist_victoria=scale(dist_victoria, center = F,scale = T))

    summary(dt4model)
```

```
##    year            vid         arm             agey     sex           elev
##  2012:1120   18      : 196   CWT:1826   Min.   :0.2   M:1759   Min.   :1137
##  2013:1187   3       : 192   SBT:1837   1st Qu.:2.4   F:1904   1st Qu.:1151
##  2014:1356   8       : 187              Median :3.5            Median :1158
##              11      : 179              Mean   :3.4            Mean   :1170
##              12      : 169              3rd Qu.:4.4            3rd Qu.:1178
##              7       : 168              Max.   :5.6            Max.   :1338
##              (Other):2572
##       tmin          prec      dist_victoria    serop_prev       selev
##  Min.   :149   Min.   : 81   Min.   :  32   Min.   :0.12   Min.   :0.97
##  1st Qu.:157   1st Qu.: 83   1st Qu.: 234   1st Qu.:0.27   1st Qu.:0.98
##  Median :159   Median : 85   Median : 713   Median :0.51   Median :0.99
##  Mean   :159   Mean   : 88   Mean   :1011   Mean   :0.50   Mean   :1.00
##  3rd Qu.:161   3rd Qu.: 94   3rd Qu.:1598   3rd Qu.:0.71   3rd Qu.:1.01
##  Max.   :161   Max.   :101   Max.   :4718   Max.   :0.87   Max.   :1.14
##
##      stmin          sprec      sdist_victoria
##  Min.   :0.94   Min.   :0.92   Min.   :0.0
##  1st Qu.:0.99   1st Qu.:0.94   1st Qu.:0.2
##  Median :1.00   Median :0.96   Median :0.5
##  Mean   :1.00   Mean   :1.00   Mean   :0.7
##  3rd Qu.:1.01   3rd Qu.:1.07   3rd Qu.:1.1
##  Max.   :1.02   Max.   :1.14   Max.   :3.3
##
```

```
    attach(dt4model)



    ##################
    # Fit a mixed-effect regression model with randomness for clusters
    ##################

    library(lme4)
    lmm <- lmer(serop_prev ~ arm + agey + sex + selev+stmin+sprec +sdist_victoria+(1 | vid),
data = dt4model,REML=F)
    summary(lmm)
```

```
## Linear mixed model fit by maximum likelihood  ['lmerMod']
## Formula:
## serop_prev ~ arm + agey + sex + selev + stmin + sprec + sdist_victoria +
##     (1 | vid)
##    Data: dt4model
##
##      AIC      BIC   logLik deviance df.resid
##   -93740   -93678    46880   -93760     3653
##
## Scaled residuals:
```

```
##       Min        1Q    Median        3Q       Max
## -1.33e-05 -1.73e-06  1.77e-06  3.08e-06  8.03e-06
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  vid      (Intercept) 1.52e-04 1.23e-02
##  Residual             3.66e-13 6.05e-07
## Number of obs: 3663, groups:  vid, 30
##
## Fixed effects:
##                 Estimate Std. Error t value
## (Intercept)     2.15e-01   1.46e-01    1.47
## armSBT          2.52e-02   4.59e-03    5.50
## agey           -2.37e-15   8.30e-09    0.00
## sexF            8.48e-14   2.01e-08    0.00
## selev           2.76e+00   1.05e-01   26.33
## stmin          -5.97e-01   1.41e-01   -4.25
## sprec          -1.78e+00   4.06e-02  -43.92
## sdist_victoria -1.91e-01   6.66e-03  -28.74
##
## Correlation of Fixed Effects:
##             (Intr) armSBT agey   sexF   selev  stmin  sprec
## armSBT      -0.012
## agey         0.000  0.000
## sexF         0.000  0.000 -0.001
## selev       -0.385  0.000  0.000  0.000
## stmin       -0.685 -0.042  0.000  0.000 -0.362
## sprec       -0.245  0.130  0.000  0.000  0.146 -0.129
## sdist_victr  0.039 -0.006  0.000  0.000 -0.762  0.627 -0.462
## optimizer (nloptwrap) convergence code: 0 (OK)
## Model failed to converge with max|grad| = 1.7506 (tol = 0.002, component 1)
## Model is nearly unidentifiable: very large eigenvalue
##  - Rescale variables?
```

```
    library(car)
    Anova(lmm) #agey and sex are not significant.
```

```
## Analysis of Deviance Table (Type II Wald chisquare tests)
##
## Response: serop_prev
##                 Chisq Df Pr(>Chisq)
## arm              30.2  1    3.9e-08 ***
## agey              0.0  1          1
## sex               0.0  1          1
## selev           693.1  1    < 2e-16 ***
## stmin            18.0  1    2.2e-05 ***
## sprec          1928.8  1    < 2e-16 ***
## sdist_victoria  826.0  1    < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
    # The output gives some measures of model fit, including AIC, BIC, log likelihood, and de
```

```
viance. Also gives an estimate of the variance explained by the random effect. The random eff
ect here is indistinguishable from 0, then the random effect may not matter so we can do a re
gular linear model instead.




    #################
    #Fit a linear regression
    #################

    #univariate linear regression
    lm_unimodel<-function(y,x,D){
    unimodel <- lm(y~x, data = D)
    s<-summary(unimodel)
    return(s)
    }
   varlist <- dt4model[,c(3,4,5,11,12,13,14)] #单变量list
   lapply(y=dt4model$serop_prev, varlist, D=dt4model, FUN=lm_unimodel)
```

```
## $arm
##
## Call:
## lm(formula = y ~ x, data = D)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.3546 -0.1900 -0.0244  0.2432  0.3750
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.45360    0.00535    84.8   <2e-16 ***
## xSBT         0.09459    0.00755    12.5   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.23 on 3661 degrees of freedom
## Multiple R-squared:  0.0411, Adjusted R-squared:  0.0409
## F-statistic:  157 on 1 and 3661 DF,  p-value: <2e-16
##
##
## $agey
##
## Call:
## lm(formula = y ~ x, data = D)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.3893 -0.2309  0.0103  0.2082  0.3781
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.51206    0.01141   44.90   <2e-16 ***
```

```
## x           -0.00326    0.00318   -1.03      0.3
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.23 on 3661 degrees of freedom
## Multiple R-squared:  0.000288,   Adjusted R-squared:  1.5e-05
## F-statistic: 1.06 on 1 and 3661 DF,  p-value: 0.304
##
##
## $sex
##
## Call:
## lm(formula = y ~ x, data = D)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.3852 -0.2347  0.0053  0.2082  0.3774
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.49558    0.00556   89.12   <2e-16 ***
## xF           0.01049    0.00771    1.36     0.17
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.23 on 3661 degrees of freedom
## Multiple R-squared:  0.000505,   Adjusted R-squared:  0.000232
## F-statistic: 1.85 on 1 and 3661 DF,  p-value: 0.174
##
##
## $selev
##
## Call:
## lm(formula = y ~ x, data = D)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.4130 -0.1812 -0.0348  0.1607  0.3852
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    2.572      0.125    20.6   <2e-16 ***
## x             -2.072      0.125   -16.6   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.22 on 3661 degrees of freedom
## Multiple R-squared:  0.0697, Adjusted R-squared:  0.0694
## F-statistic:  274 on 1 and 3661 DF,  p-value: <2e-16
##
##
## $stmin
##
## Call:
```

```
## lm(formula = y ~ x, data = D)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.3869 -0.1701  0.0099  0.0881  0.3546
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -9.200      0.202   -45.5   <2e-16 ***
## x              9.704      0.202    48.0   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.18 on 3661 degrees of freedom
## Multiple R-squared:  0.386,  Adjusted R-squared:  0.386
## F-statistic: 2.3e+03 on 1 and 3661 DF,  p-value: <2e-16
##
##
## $sprec
##
## Call:
## lm(formula = y ~ x, data = D)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.4339 -0.1173  0.0175  0.1358  0.1998
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.1214     0.0369    84.5   <2e-16 ***
## x            -2.6267     0.0369   -71.1   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.15 on 3661 degrees of freedom
## Multiple R-squared:  0.58,   Adjusted R-squared:  0.58
## F-statistic: 5.06e+03 on 1 and 3661 DF,  p-value: <2e-16
##
##
## $sdist_victoria
##
## Call:
## lm(formula = y ~ x, data = D)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.3912 -0.0698 -0.0242  0.0969  0.3267
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.65437    0.00409   160.1   <2e-16 ***
## x           -0.21727    0.00409   -53.1   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 0.17 on 3661 degrees of freedom
## Multiple R-squared:  0.435,  Adjusted R-squared:  0.435
## F-statistic: 2.82e+03 on 1 and 3661 DF,  p-value: <2e-16
```

```
   #Based on the univariate linear regression output, agey and sex are not significant.

    #Fit a linear regression model
    lmfull<-lm(serop_prev ~ arm + agey + sex +selev+stmin+sprec+sdist_victoria, data=dt4model
)
    summary(lmfull) #agey, sex,selev are not significant.
```

```
##
## Call:
## lm(formula = serop_prev ~ arm + agey + sex + selev + stmin +
##     sprec + sdist_victoria, data = dt4model)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.3562 -0.0610  0.0059  0.0853  0.2087
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    11.750846   1.007807   11.66  < 2e-16 ***
## armSBT          0.033855   0.004365    7.76  1.1e-14 ***
## agey            0.000135   0.001769    0.08     0.94
## sexF            0.005643   0.004293    1.31     0.19
## selev          -0.093972   0.227348   -0.41     0.68
## stmin          -8.216082   0.703104  -11.69  < 2e-16 ***
## sprec          -2.837619   0.107402  -26.42  < 2e-16 ***
## sdist_victoria -0.186308   0.005603  -33.25  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.13 on 3655 degrees of freedom
## Multiple R-squared:  0.692,  Adjusted R-squared:  0.691
## F-statistic: 1.17e+03 on 7 and 3655 DF,  p-value: <2e-16
```

```
    lm1<-lm(serop_prev ~ arm +stmin+sprec+sdist_victoria, data=dt4model)
    summary(lm1)
```

```
##
## Call:
## lm(formula = serop_prev ~ arm + stmin + sprec + sdist_victoria,
##     data = dt4model)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.3529 -0.0574  0.0033  0.0827  0.2049
##
## Coefficients:
```

```
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    11.37807    0.38237   29.76   <2e-16 ***
## armSBT          0.03384    0.00436    7.77    1e-14 ***
## stmin          -7.97084    0.34008  -23.44   <2e-16 ***
## sprec          -2.79979    0.05556  -50.40   <2e-16 ***
## sdist_victoria -0.18724    0.00531  -35.24   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.13 on 3658 degrees of freedom
## Multiple R-squared:  0.691,  Adjusted R-squared:  0.691
## F-statistic: 2.05e+03 on 4 and 3658 DF,  p-value: <2e-16
```

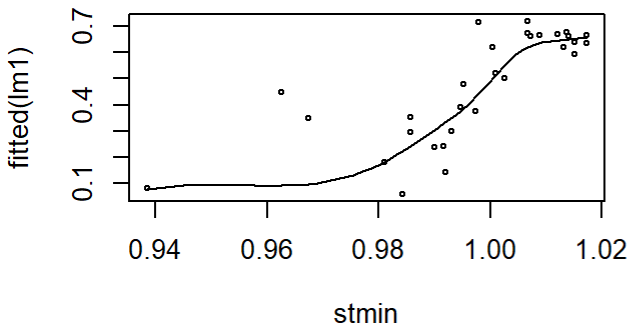
```
    #The standard deviance of stmin is much larger than those of other variables. Could have collinearity issue.

    # Model selection
    lmnull<-lm(serop_prev ~ 1, data=dt4model)
    stepAIC(lmnull,direction="both",scope=list(upper=lmfull,lower=lmnull))
```

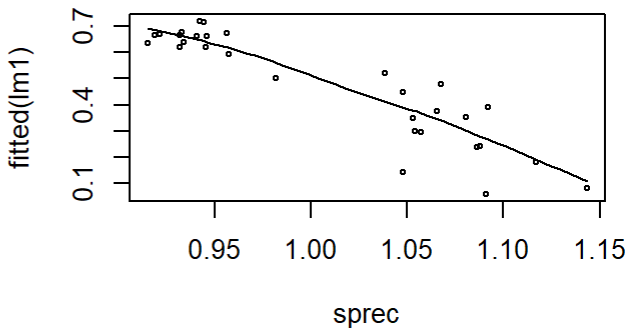
```
## Start:  AIC=-10663
## serop_prev ~ 1
##
##                  Df Sum of Sq   RSS    AIC
## + sprec           1     115.6  83.6 -13841
## + sdist_victoria  1      86.7 112.5 -12755
## + stmin           1      76.9 122.3 -12447
## + selev           1      13.9 185.3 -10926
## + arm             1       8.2 191.0 -10815
## <none>                        199.2 -10663
## + sex             1       0.1 199.1 -10663
## + agey            1       0.1 199.2 -10662
##
## Step:  AIC=-13841
## serop_prev ~ sprec
##
##                  Df Sum of Sq   RSS    AIC
## + sdist_victoria  1      11.4  72.2 -14375
## + arm             1       1.3  82.4 -13894
## + selev           1       0.3  83.3 -13853
## + sex             1       0.1  83.5 -13844
## <none>                         83.6 -13841
## + stmin           1       0.0  83.6 -13839
## + agey            1       0.0  83.6 -13839
## - sprec           1     115.6 199.2 -10663
##
## Step:  AIC=-14375
## serop_prev ~ sprec + sdist_victoria
##
##                  Df Sum of Sq   RSS    AIC
## + stmin           1       9.8  62.5 -14904
## + selev           1       7.5  64.7 -14775
## + arm             1       1.5  70.7 -14452
```

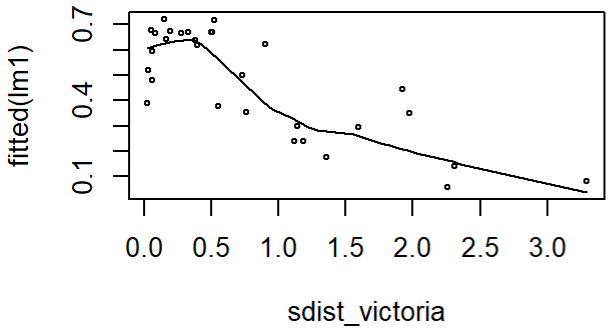
```
## + sex             1       0.1  72.2 -14376
## <none>                        72.2 -14375
## + agey            1       0.0  72.2 -14373
## - sdist_victoria  1      11.4  83.6 -13841
## - sprec           1      40.2 112.5 -12755
##
## Step:  AIC=-14904
## serop_prev ~ sprec + sdist_victoria + stmin
##
##                  Df Sum of Sq   RSS    AIC
## + arm             1       1.0  61.5 -14962
## <none>                        62.5 -14904
## + sex             1       0.0  62.5 -14904
## + agey            1       0.0  62.5 -14902
## + selev           1       0.0  62.5 -14902
## - stmin           1       9.8  72.2 -14375
## - sdist_victoria  1      21.1  83.6 -13839
## - sprec           1      46.0 108.5 -12885
##
## Step:  AIC=-14962
## serop_prev ~ sprec + sdist_victoria + stmin + arm
##
##                  Df Sum of Sq   RSS    AIC
## <none>                        61.5 -14962
## + sex             1       0.0  61.4 -14962
## + selev           1       0.0  61.5 -14960
## + agey            1       0.0  61.5 -14960
## - arm             1       1.0  62.5 -14904
## - stmin           1       9.2  70.7 -14452
## - sdist_victoria  1      20.9  82.4 -13893
## - sprec           1      42.7 104.2 -13033
```

```
##
## Call:
## lm(formula = serop_prev ~ sprec + sdist_victoria + stmin + arm,
##     data = dt4model)
##
## Coefficients:
##    (Intercept)           sprec  sdist_victoria           stmin          armSBT
##        11.3781         -2.7998         -0.1872         -7.9708          0.0338
```

```
    #STEPAIC model:exactly the same as lm1.


    #Check model linear assumptions.
    par(mfrow=c(2,2))
    scatter.smooth(stmin, fitted(lm1), cex=0.5)
    scatter.smooth(sprec,fitted(lm1), cex=0.5)
    scatter.smooth(sdist_victoria, fitted(lm1),cex=0.5)
    #Basically all follow linear assumption; stmin and sdist_victoria show a bit flat on the
one end, due to some outliers in that region.
    par(mfrow=c(1,1))
```

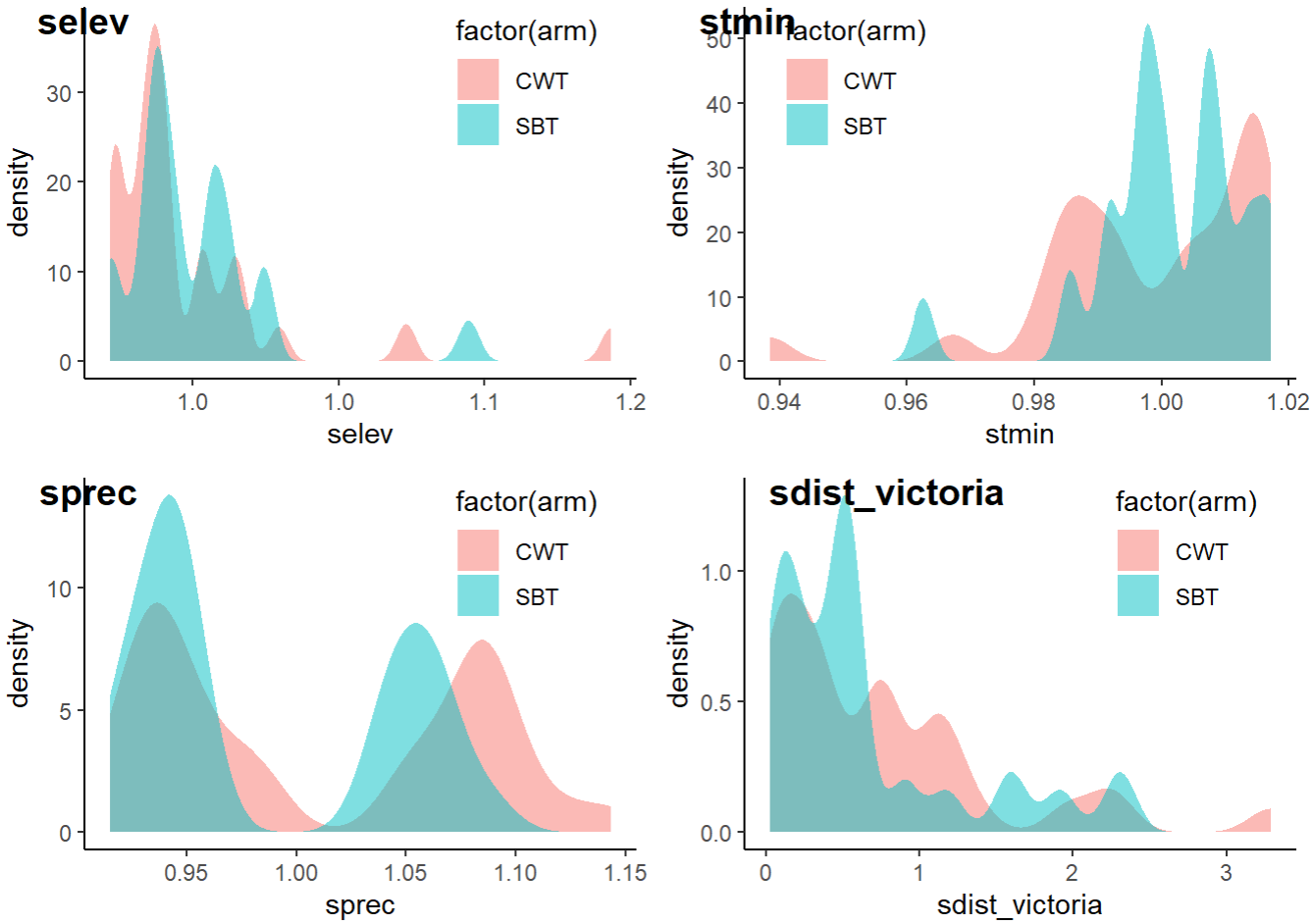




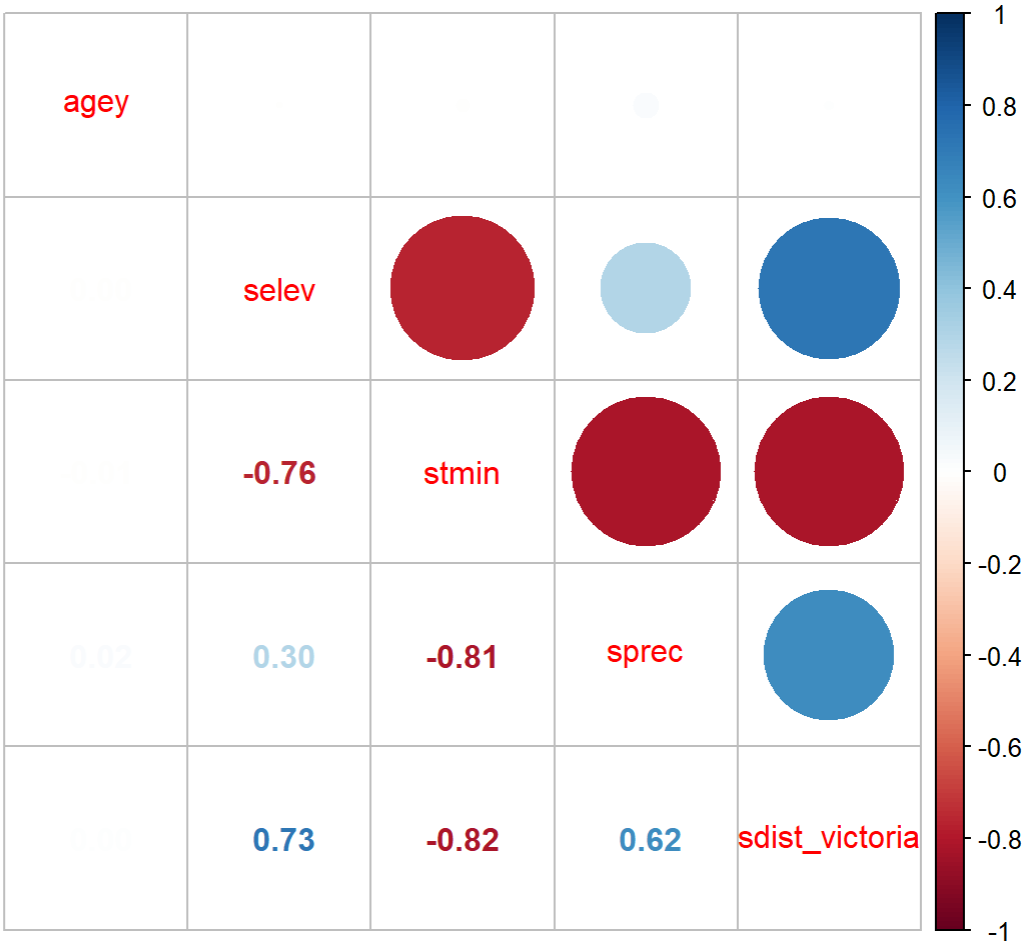



```
# density plots of the geographical variables, broken down by 2 groups.
library(cowplot)
library(ggplot2)

p1<-ggplot(complete, aes(selev, fill=factor(arm))) +
geom_density(alpha=.5,color=NA)+
theme_classic()+
theme(legend.position = c(.8, .8))

p2<-ggplot(complete, aes(stmin, fill=factor(arm))) +
geom_density(alpha=.5,color=NA)+
theme_classic()+
theme(legend.position = c(.2, .8))

p3<-ggplot(complete, aes(sprec, fill=factor(arm))) +
geom_density(alpha=.5,color=NA)+
theme_classic()+
theme(legend.position = c(.8, .8))

p4<-ggplot(complete, aes(sdist_victoria, fill=factor(arm))) +
geom_density(alpha=.5,color=NA)+
theme_classic()+
theme(legend.position = c(.8, .8))

plot_grid(p1,p2,p3,p4, nrow=2,ncol=2, labels=c('selev','stmin','sprec','sdist_victoria'),
```

```
align=c('v','h'))
```




```
    #from the density plot, the distribution of selev is not quite different between 2 groups
.


     #correlation plot to check collinearity issue
    library(corrplot)
    rr <- cor(dt4model[ ,c(4,11,12,13,14)])
    corrplot.mixed(rr)
```




```
    #vif
    library(car)
    vif(lm(serop_prev ~ arm +stmin+sprec + sdist_victoria, data=dt4model))
```

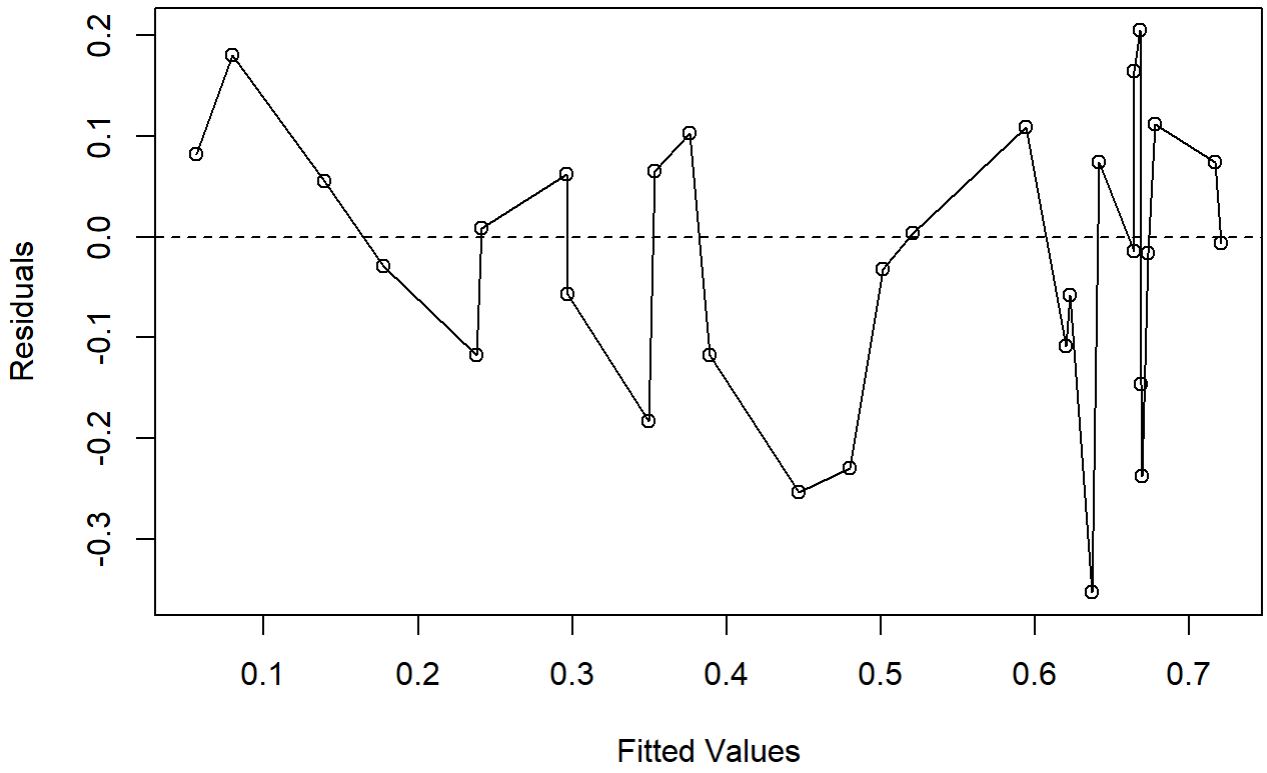
```
##            arm          stmin          sprec sdist_victoria
##            1.0            5.6            3.1            3.1
```

```
    #The four geographical variables are highly correlated - stmin has strong relationships w
ith all three other terms, would consider not to keep all of them in the model.


    #drop stmin and build a model2 with fewer variables.
    lm2<-lm(serop_prev ~ arm +sprec + sdist_victoria, data=dt4model)
    summary(lm2)
```

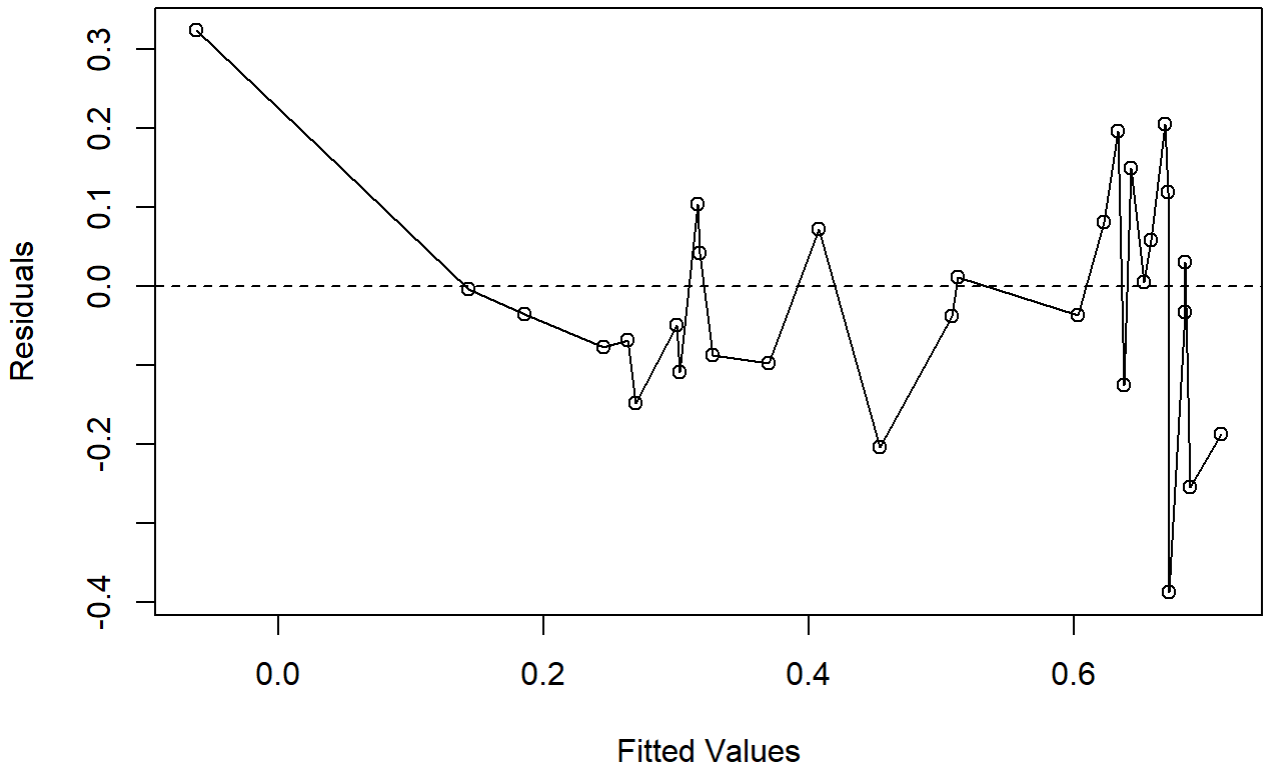
```
##
## Call:
## lm(formula = serop_prev ~ arm + sprec + sdist_victoria, data = dt4model)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.3877 -0.0873  0.0045  0.1034  0.3230
##
## Coefficients:
```

```
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     2.46406    0.04242   58.09   <2e-16 ***
## armSBT          0.04151    0.00466    8.91   <2e-16 ***
## sprec          -1.91679    0.04378  -43.78   <2e-16 ***
## sdist_victoria -0.10161    0.00414  -24.57   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.14 on 3659 degrees of freedom
## Multiple R-squared:  0.645,  Adjusted R-squared:  0.645
## F-statistic: 2.22e+03 on 3 and 3659 DF,  p-value: <2e-16
```

```
    # plot residuals against the fitted values
    plot(fitted(lm1), residuals(lm1), xlab="Fitted Values", ylab="Residuals")
    lines(smooth.spline(fitted(lm1), residuals(lm1)))
    abline(h=0, lty=2)
```




```
    plot(fitted(lm2), residuals(lm2), xlab="Fitted Values", ylab="Residuals")
    lines(smooth.spline(fitted(lm2), residuals(lm2)))
    abline(h=0, lty=2)
```




```
    # we expect to see the random scatter. If the scatter is not random that means there's so
me variation in the data that has not been explained.A dashed horizontal line representing 0:
an average of 0 deviaiton from the best fit line; a solid line represents the residual deviat
ion from the best fit line. Ideally, it will overlay the dashed line.
    #The plots indicate that neither model is a good fit for the data. In both plots, we can
observe a pattern in the distribution of residuals, indicates that the variance is not homoge
neous. This can be explained by the presence of clusters, as each cluster may have a differen
t variance structure.



    ##################
    #Fit a GEE model for a marginal analysis
    ##################
    # GEE:generalized estimate equation: to try different variance structures
    library(geepack)

    #First we try independence structure, i.e.the linear regression iid variance structure.
    gee1<-geeglm(serop_prev ~ arm +sprec + sdist_victoria, data=dt4model, id=vid, family = ga
ussian, corstr="independence",std.err = "san.se")
    summary(gee1)
```

```
##
## Call:
## geeglm(formula = serop_prev ~ arm + sprec + sdist_victoria, family = gaussian,
```

```
##     data = dt4model, id = vid, corstr = "independence", std.err = "san.se")
##
##  Coefficients:
##                Estimate Std.err  Wald Pr(>|W|)
## (Intercept)      2.4641  0.3216 58.71  1.8e-14 ***
## armSBT           0.0415  0.0325  1.63      0.2
## sprec           -1.9168  0.3177 36.39  1.6e-09 ***
## sdist_victoria  -0.1016  0.0230 19.45  1.0e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation structure = independence
## Estimated Scale Parameters:
##
##             Estimate Std.err
## (Intercept)   0.0193 0.00313
## Number of clusters:   90  Maximum cluster size: 75
```

```
    #Based on the previous exploratory analysis, we use a linear model for the mean for both
groups;
    # the coefficient estimations are the samle as of lm2, due to the independence variance s
tructure.


    #now fit a proper model
    #with compound symmetric (constant) correlation structure for depdendence in clusters.
    # Assume same for 2 groups.
    gee2<-geeglm(serop_prev ~ arm +sprec + sdist_victoria, data=dt4model, id=vid, family = ga
ussian, corstr="exchangeable", std.err = "san.se")
    summary(gee2)
```

```
##
## Call:
## geeglm(formula = serop_prev ~ arm + sprec + sdist_victoria, family = gaussian,
##     data = dt4model, id = vid, corstr = "exchangeable", std.err = "san.se")
##
##  Coefficients:
##                Estimate Std.err  Wald Pr(>|W|)
## (Intercept)      2.4225  0.2932 68.27  < 2e-16 ***
## armSBT           0.0271  0.0309  0.77   0.3813
## sprec           -1.9098  0.2857 44.67  2.3e-11 ***
## sdist_victoria  -0.0736  0.0239  9.50   0.0021 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation structure = exchangeable
## Estimated Scale Parameters:
##
##             Estimate Std.err
## (Intercept)   0.0203  0.0033
##   Link = identity
##
## Estimated Correlation Parameters:
```

```
##       Estimate Std.err
## alpha     1.06  0.0408
## Number of clusters:   90  Maximum cluster size: 75
```

```
    # now try an autoregressive correlation AR1
    gee3<-geeglm(serop_prev ~ arm +sprec + sdist_victoria, data=dt4model, id=vid, family = ga
ussian, corstr="ar1", std.err = "san.se")
    summary(gee3)
```

```
##
## Call:
## geeglm(formula = serop_prev ~ arm + sprec + sdist_victoria, family = gaussian,
##     data = dt4model, id = vid, corstr = "ar1", std.err = "san.se")
##
##  Coefficients:
##                Estimate Std.err  Wald Pr(>|W|)
## (Intercept)      2.4123  0.2916 68.45  < 2e-16 ***
## armSBT           0.0259  0.0309  0.70   0.4012
## sprec           -1.9032  0.2836 45.02  1.9e-11 ***
## sdist_victoria  -0.0715  0.0238  8.99   0.0027 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation structure = ar1
## Estimated Scale Parameters:
##
##             Estimate Std.err
## (Intercept)   0.0205 0.00334
##   Link = identity
##
## Estimated Correlation Parameters:
##       Estimate Std.err
## alpha        1 0.00219
## Number of clusters:   90  Maximum cluster size: 75
```

```
    #Finally, try a unstructured correlation/a non-parametric form: heterogeneous variance
    #may take longer to run, due to large number of villages.
    # gee4<-geeglm(serop_prev ~ arm +sprec + sdist_victoria, data=dt4model, id=vid, family =
gaussian, corstr="unstructured", std.err = "san.se")
    # summary(gee4)


    #Since that all the gee models shows no significant effect of arm adjusted on other varia
bles, try to fit a univarite model with structured variance.
    gee0<-geeglm(serop_prev ~ arm, data=dt4model, id=vid, family = gaussian, corstr="exchange
able", std.err = "san.se")
    summary(gee0)
```

```
##
## Call:
## geeglm(formula = serop_prev ~ arm, family = gaussian, data = dt4model,
```

```
##     id = vid, corstr = "exchangeable", std.err = "san.se")
##
##  Coefficients:
##             Estimate Std.err   Wald Pr(>|W|)
## (Intercept)   0.4044  0.0362 125.13   <2e-16 ***
## armSBT        0.0925  0.0478   3.74    0.053 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation structure = exchangeable
## Estimated Scale Parameters:
##
##             Estimate Std.err
## (Intercept)   0.0547 0.00598
##   Link = identity
##
## Estimated Correlation Parameters:
##       Estimate Std.err
## alpha     1.06  0.0441
## Number of clusters:   90  Maximum cluster size: 75
```

```
    #Now without other variables in the model, the P-value of armSBT alone is 0.053, close to o.o5, but still not significant.

    #We can see that different correlation structures produce different results. We can select the covariance structure which is most appropriate for the model by comparing the AIC or BIC values for each model and select the one with the smallest value.

    #Provide estimates of effect, 95% confidence intervals, and a P-value for the difference. Summarize your results in a formatted table.
    library(doBy)
    est=esticon(gee2,diag(4))
    sum_table<-est[, -c(5,6)]
    rownames(sum_table) <-c("Intercept", "ArmSBT","Sprec", "Sdist_victoria")
    colnames(sum_table) <- c("Estimate", "Standard error","Statistic", "P-value","Lower","Upper")
    head(sum_table,6)
```

```
##                 Estimate Standard error Statistic   P-value     Lower Upper
## Intercept       2.42e+00       2.93e-01  6.83e+01  1.11e-16  1.85e+00  3.00
## ArmSBT          2.71e-02       3.09e-02  7.66e-01  3.81e-01 -3.35e-02  0.09
## Sprec          -1.91e+00       2.86e-01  4.47e+01  2.33e-11 -2.47e+00 -1.35
## Sdist_victoria -7.36e-02       2.39e-02  9.50e+00  2.05e-03 -1.20e-01 -0.03
```

```
    #Provide a brief interpretation of the results.
   # After analyzing the outputs from the GEE models, it was found that the estimated effect of the "arm" variable was 0.04. This means that, under the same conditions of "pre" and "dist_victoria", the prevalence is expected to increase by 0.04 when transitioning from the CWT group to the SBT group. However, since the P-value is greater than 0.05, there is no statistically significant difference in the prevalence between the CWT and SBT groups over the two years, whether the other variables are adjusted or not in the model. It's important to note that this conclusion differs from that of the general linear regression model, as the GEE model
```

```
takes into account the cluster effect while modeling the data.
```

2. If you think that comparing groups using a relative measure of effect would be better, such as the prevalence ratio or odds ratio, then briefly justify your reasoning and approach.

Research papers have suggested using different models for analyzing data in studies that focus on changes within participants over time or clusters of participants. However, these models may not work well in certain situations where the data is binary. When the data is Gaussian, the results from both models are similar, but with binary data, the results from the marginal model can be smaller and have a different interpretation.

Please see: @ (David M. Murray, PhD, Sherri P. Varnell, PhD, MS, and Jonathan L. Blitstein, MS: Design and Analysis of Group-Randomized Trials: A Review of Recent Methodological Developments: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1448268/)

#5 Bonus Challenge (model validation)

Conditional on enrollment into a trial, a randomized controlled trial has one source of random variation: the treatment assignment. An approach to exact inference in a trial is to compare groups using a permutation test, where the treatment assignment is re-randomized across many (sometime all) permutations and a test statistic is computed in each permuted dataset. The distribution of the test statistic over the permutations defines its null distribution, which enables exact inference.

Most trials at Proctor rely on permutation tests for our primary inference. For one of the outcomes above, estimate the permutation P-value for differences between groups, assuming that the only random variation in the trial is the community-level treatment assignment (which should be approximately true, by design!). You can permute any test statistic you want. How does your inference compare with your results from the previous section?

```
#Idea: Boostrap method for any test statistic from the previous sections.
#Treat the sample we have as the real population, randomly select a equal number of observati
ons from the population and  use as a new sample, model on that sample to get a test statist
ic; repeated this process for many times(like 100 times) and get many statistics (like 100 st
atistics) to form a distribution of this statistics. Calculate the P value of the statistic t
hat we got from last section.

    #using BOOSTRAP resampling method to build up a distribution for the test statistic of th
e coefficient estimation of arm.
    gl<-geeglm(serop_prev ~ arm, data=dt4model, id=vid, family = gaussian, corstr="exchangeab
le", std.err = "san.se")
    a<--as.numeric(unlist(summary(gl)[6])[6])


    #a boostrap function
    a<-NULL
    statis.fn <- function(data,number,index){
      set.seed(1)
      for (i in index) {
      library(dplyr)
      dt <- sample_n(data, number)
      gl<-geeglm(serop_prev ~ arm, data=dt, id=vid, family = gaussian, corstr="exchangeable",
std.err = "san.se")
      a<--as.numeric(unlist(summary(gl)[6])[6])
      a <- c(a, i)
      }
      return(a)
    }
```

```
   #perform the bootstrap analysis with 100 resamples
     bootstrap_stat<-statis.fn(dt4model,100,1:10)
     # hist(bootstrap_stat)

   #calculate the 95% confidence interval for the mean estimate
     lower_ci <- quantile(bootstrap_stat, 0.025)
     upper_ci <- quantile(bootstrap_stat, 0.975)

     c(lower_ci,upper_ci)
```

```
##  2.5% 97.5%
## -1.99  9.69
```

```
   # Other example: compare std.error using boostrap from boot package.
    library(boot)
    boot.fn=function(data,index)
    coefficients(lm(serop_prev ~ arm, data=data, id=vid, subset = index))
    set.seed(1)
    boot(dt4model, boot.fn, 100)
```

```
##
## ORDINARY NONPARAMETRIC BOOTSTRAP
##
##
## Call:
## boot(data = dt4model, statistic = boot.fn, R = 100)
##
##
## Bootstrap Statistics :
##     original    bias    std. error
## t1*   0.4536 -0.000653     0.00564
## t2*   0.0946  0.001060     0.00689
```

```
   #compared with estimates of armSBT from formulas directly.
   summary(lm(serop_prev ~ arm, data=dt4model))$coef
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.4536    0.00535    84.9 0.00e+00
## armSBT        0.0946    0.00755    12.5 2.65e-35
```

```
  detach(complete)
```