

Analysis of synthetic data using WCLS

Tianchen Qian

May 30, 2021

The data here is a synthetic data set mimicking some features of the HeartSteps data set.

1. Preparation

We load the already generated synthetic data set and load some packages and functions required to carry out the WCLS analysis.

```
rm(list = ls())

library(tidyverse)
library(xtable)
library(geepack)
source("xgeepack.R")
source("estimate.R")

synthetic_data <- read.csv("synthetic_data_37subject_210time.csv")

summary(synthetic_data)
```

##	userid	decision.index.nogap	study.day.nogap	jbsteps30.log
##	Min. : 1	Min. : 1.0	Min. : 0.0	Min. : -0.6931
##	1st Qu.: 10	1st Qu.: 53.0	1st Qu.: 10.0	1st Qu.: 0.5519
##	Median : 19	Median : 105.5	Median : 20.5	Median : 2.5680
##	Mean : 19	Mean : 105.5	Mean : 20.5	Mean : 2.7204
##	3rd Qu.: 28	3rd Qu.: 158.0	3rd Qu.: 31.0	3rd Qu.: 4.4883
##	Max. : 37	Max. : 210.0	Max. : 41.0	Max. : 8.5000

##	jbsteps30.log.lag1	jbsteps30pre.log	location.homework	send
##	Min. : -0.6931	Min. : -0.6931	Min. : 0.0000	Min. : 0.000
##	1st Qu.: 0.5156	1st Qu.: -0.6931	1st Qu.: 0.0000	1st Qu.: 0.000
##	Median : 2.5514	Median : 3.0074	Median : 0.0000	Median : 0.000
##	Mean : 2.7068	Mean : 2.1025	Mean : 0.3683	Mean : 0.497
##	3rd Qu.: 4.4813	3rd Qu.: 4.0007	3rd Qu.: 1.0000	3rd Qu.: 1.000
##	Max. : 8.5000	Max. : 5.3293	Max. : 1.0000	Max. : 1.000

##	avail
##	Min. : 0.0000
##	1st Qu.: 1.0000
##	Median : 1.0000
##	Mean : 0.8049
##	3rd Qu.: 1.0000
##	Max. : 1.0000

The variable names are kept consistent with the original HeartSteps data set and the analysis code for that data, the analysis result of which is included in the main manuscript. Below are some explanation for each of

the variables:

- `userid`: id of a user (ranging from 1 to 37)
- `decision.index.nogap`: decision point index for each user (ranging from 1 to 210)
- `study.day.nogap`: day in the study for each user (ranging from 0 to 41)
- `jbsteps30.log`: log-transformed 30-minute step count following each decision point (it is called `jbsteps` because in HeartSteps the step count was measured by Jawbone tracker)
- `jbsteps30.log.lag1`: log-transformed 30-minute step count following the previous decision point; i.e., a lagged version of `jbsteps30.log`
- `jbsteps30pre.log`: log-transformed step count in the 30-minute window *preceding* each decision point
- `location.homework`: an indicator of whether the user is currently at home/work (1) or other places (0)
- `send`: treatment indicator, whether an activity suggestion was sent at the decision point
- `avail`: availability indicator, whether the person is available at the decision point

We create two additional variables to be used in the WCLS regression below.

```
synthetic_data$"(Intercept)" <- 1
synthetic_data$"I(send - 0.6)" <- synthetic_data$send - 0.6
```

2. Using WCLS to analyze the data

We use the Weighted and Centered Least Squares (WCLS) estimator to analyze the data.

2.1. Marginal Effect

In this analysis, we aim to answer the question: “What is the effect of delivering activity suggestions on individuals’ subsequent 30-minute step counts?”

```
xmat <- synthetic_data %>%
  transmute("(Intercept)" = .$(Intercept)",
            "jbsteps30pre.log" = .$(jbsteps30pre.log)",
            "I(send - 0.6)" = .$(I(send - 0.6))")

fit_model1 <- geese.glm(x = as.matrix(xmat),
                      y = synthetic_data$jbsteps30.log,
                      w = synthetic_data$avail,
                      id = as.factor(synthetic_data$user),
                      family = gaussian(), corstr = "independence")

estimate(fit_model1)
```

##	Estimate	95% LCL	95% UCL	SE	Hotelling	df1	df2
## (Intercept)	2.01e+00	1.92e+00	2.10e+00	4.57e-02	1.94e+03	1.00e+00	34
## jbsteps30pre.log	3.40e-01	3.00e-01	3.80e-01	1.97e-02	2.98e+02	1.00e+00	34
## I(send - 0.6)	1.57e-01	3.10e-02	2.84e-01	6.22e-02	6.40e+00	1.00e+00	34
##	p-value						
## (Intercept)	<1e-04	***					
## jbsteps30pre.log	<1e-04	***					
## I(send - 0.6)	0.0162	*					
## ---							
## Signif. codes:	0 '***'	0.001 '**'	0.01 '*'	0.05 '.'	0.1 ' '	1	

2.2. Effect Change Over Time

In this analysis, we aim to answer the question: “How does the effect of activity suggestions change with each additional day in the study?”

```
xmat <- synthetic_data %>%
  transmute("(Intercept)" = .$(Intercept)",
            "jbsteps30pre.log" = .$(jbsteps30pre.log",
            "study.day.nogap" = .$(study.day.nogap",
            "I(send - 0.6)" = .$(I(send - 0.6)",
            "I(send - 0.6):study.day.nogap" =
              .$(I(send - 0.6)" * .$(study.day.nogap")

fit_model2 <- geese.glm(x = as.matrix(xmat),
                      y = synthetic_data$jbsteps30.log,
                      w = synthetic_data$avail,
                      id = as.factor(synthetic_data$user),
                      family = gaussian(), corstr = "independence")

estimate(fit_model2)
```

##	Estimate	95% LCL	95% UCL	SE	Hotelling
## (Intercept)	2.18752	2.04533	2.32971	0.06981	982.04017
## jbsteps30pre.log	0.33967	0.30003	0.37930	0.01946	304.75364
## study.day.nogap	-0.00852	-0.01398	-0.00305	0.00268	10.08379
## I(send - 0.6)	0.64860	0.43050	0.86670	0.10707	36.69331
## I(send - 0.6):study.day.nogap	-0.02374	-0.03279	-0.01469	0.00444	28.55599

```
##
##              df1 df2 p-value
## (Intercept)    1.00000 32 <1e-04 ***
## jbsteps30pre.log    1.00000 32 <1e-04 ***
## study.day.nogap    1.00000 32 0.0033 **
## I(send - 0.6)    1.00000 32 <1e-04 ***
## I(send - 0.6):study.day.nogap    1.00000 32 <1e-04 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We make the plot of the estimated effect over time along with its pointwise 95% confidence interval.

```
beta_index <- 4:5

beta_hat <- coef(fit_model2)[beta_index]
vcov <- fit_model2$geese$vbeta[beta_index, beta_index]

newdta <- df_tx <- data.frame(Intercept = 1, study.day.nogap = 0:41)
df_tx$treatment_effect <- as.matrix(newdta) %*% beta_hat
df_tx$tx_se <- NA
for (i in 1:nrow(df_tx)) {
  f_t <- as.numeric(newdta[i, ]) # feature
  df_tx$tx_se[i] <- sqrt(t(f_t) %*% vcov %*% f_t)
}
df_tx$left_ci <- df_tx$treatment_effect - 1.96 * df_tx$tx_se
df_tx$right_ci <- df_tx$treatment_effect + 1.96 * df_tx$tx_se

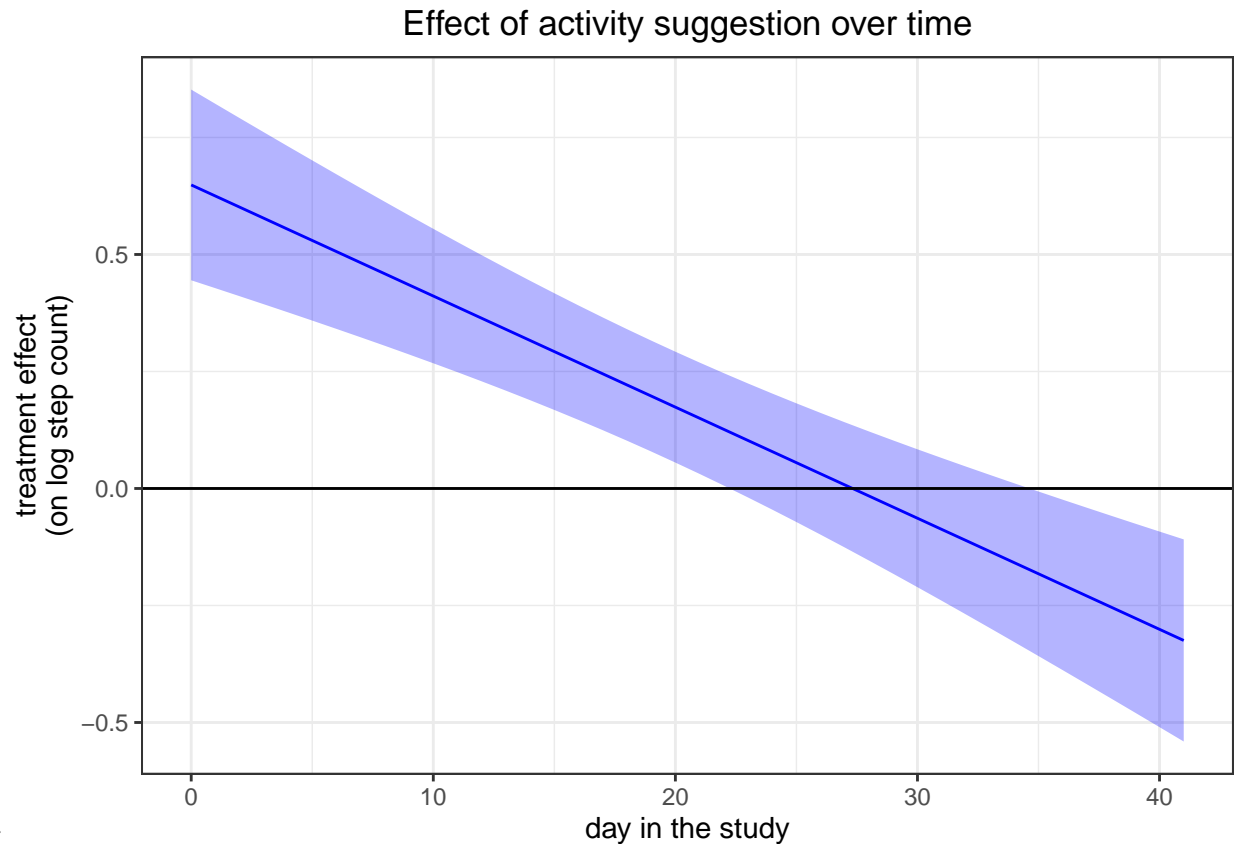
df_tx_linear <- df_tx

ggplot(df_tx) +
```

```

geom_line(aes(x = study.day.nogap, y = treatment_effect), color = "blue") +
geom_ribbon(aes(ymin = left_ci, ymax = right_ci, x = study.day.nogap),
          alpha = 0.3, fill = "blue") +
geom_hline(yintercept = 0, color = "black") +
xlab(label = "day in the study") +
ylab(label = "treatment effect\n(on log step count)") +
ggtitle(paste0("Effect of activity suggestion over time")) +
theme_bw() +
theme(plot.title = element_text(hjust = 0.5))

```



plot-1.pdf

2.3. Effect Moderated by Outcome at Previous Time Point

In this analysis, we aim to answer the question: “How does the effect of activity suggestions depend on the logged step count at previous decision point?”

```

xmat <- synthetic_data %>%
  transmute("(Intercept)" = .$(Intercept)",
            "jbsteps30pre.log" = .$(jbsteps30pre.log",
            "jbsteps30.log.lag1" = .$(jbsteps30.log.lag1",
            "location.homework" = .$(location.homework",
            "I(send - 0.6)" = .$(I(send.active - 0.6)",
            "I(send - 0.6):jbsteps30.log.lag1" =
              .$(I(send - 0.6)" * .$(jbsteps30.log.lag1")

fit_model3 <- geese.glm(x = as.matrix(xmat),
                      y = synthetic_data$jbsteps30.log,

```

```

w = synthetic_data$avail,
id = as.factor(synthetic_data$user),
family = gaussian(), corstr = "independence")
estimate(fit_model3)

```

```

##              Estimate    95% LCL    95% UCL      SE
## (Intercept)    1.85e+00  1.74e+00  1.96e+00  5.34e-02
## jbsteps30pre.log  3.41e-01  3.01e-01  3.81e-01  1.97e-02
## jbsteps30.log.lag1 3.87e-02  1.55e-02  6.19e-02  1.14e-02
## location.homework 1.51e-01  3.93e-02  2.63e-01  5.48e-02
## I(send - 0.6):jbsteps30.log.lag1 2.83e-02 -1.66e-03  5.82e-02  1.47e-02
##              Hotelling      df1 df2 p-value
## (Intercept)    1.20e+03  1.00e+00  32 < 1e-04 ***
## jbsteps30pre.log  3.01e+02  1.00e+00  32 < 1e-04 ***
## jbsteps30.log.lag1 1.16e+01  1.00e+00  32 0.00183 **
## location.homework  7.58e+00  1.00e+00  32 0.00964 **
## I(send - 0.6):jbsteps30.log.lag1 3.70e+00  1.00e+00  32 0.06326 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```