Multiple Imputation Edge Cases

2022-07-22, Leo Murao Watson

Special Cases where Listwise Deletion is Preferred over Multiple Imputation

1) Exclusively Missing data in Response Y

- Let Y = Ozone, X_1 = Wind, X_2 = Temp, X_3 = Month, X_4 = Day
- Will compare Missing Imputation and Listwise Deletion as missing data methods.

Missing Imputation

```
simulate_MI <- function(runs = 10) {
  airquality_processed <- airquality %>% select(Ozone, Wind, Temp, Month, Day)
  res <- array(NA, dim = c(5, runs, 3))
  dimnames(res) <- list(c("Intercept", "Wind", "Temp", "Month", "Day"),
                        as.character(1:runs), c("estimate", "2.5%", "97.5%"))
  for (run in 1:runs){
    imp_MI <- mice(airquality_processed, print = FALSE)
    fit <- with(imp_MI, lm(Ozone ~ Wind + Temp + Month + Day))
    tab <- summary(pool(fit), "all", conf.int = TRUE)
    res[1, run, ] <- as.numeric(tab[1, c("estimate", "2.5 %", "97.5 %")])
    res[2, run, ] <- as.numeric(tab[2, c("estimate", "2.5 %", "97.5 %")])
    res[3, run, ] <- as.numeric(tab[3, c("estimate", "2.5 %", "97.5 %")])
    res[4, run, ] <- as.numeric(tab[4, c("estimate", "2.5 %", "97.5 %")])
    res[5, run, ] <- as.numeric(tab[5, c("estimate", "2.5 %", "97.5 %")])
  }
  res
}
```

```
# Measure time taken for simuluating multiple imputation
start_time <- Sys.time()
res_MI <- simulate_MI(100)
end_time <- Sys.time()
end_time - start_time
```

```
## Time difference of 11.75224 secs
```

```
# Obtain confidence intervals & estimates for all coefficients, intercept.
apply(res_MI, c(1, 3), mean, na.rm = TRUE)
```

```
##               estimate         2.5%      97.5%
## Intercept -60.8498437 -107.7961034 -13.903584
## Wind       -3.1210129   -4.4821762  -1.759850
## Temp        2.0016656    1.4700953   2.533236
## Month      -3.6346773   -6.7098267  -0.559528
## Day         0.2432321   -0.2233079   0.709772
```

Listwise Deletion

```
simulate_LD <- function(runs = 10){
  lw_airquality <- airquality %>% select(Ozone, Wind, Temp, Month, Day) %>%
    filter(!is.na(Ozone))
  res <- array(NA, dim = c(5, runs, 3))
  dimnames(res) <- list(c("Intercept", "Wind", "Temp", "Month", "Day"),
                        as.character(1:runs), c("estimate", "2.5%", "97.5%"))
  # Loop over each iteration
  for (run in 1:runs){
    fit <- with(lw_airquality, lm(Ozone ~ Wind + Temp + Month + Day))
    # loop over each variable
    for (var in 1:5){
    edges <- as.numeric((confint(fit)[var,]))
    mid <- as.numeric(fit$coefficients)[var]
    interval <- c(edges[1], mid, edges[2])
    res[var, run, ] <- interval
    }
  }
  res
}
```

```
# REMARK: no randomness in the listwise deletion process so deterministic
# and hence doing 1 sim <=> 1000 sims. This will affect running time.
# In order to account for this, this code chunk multiplies time for
# single occurence of listwise deletion and multiplies
# by # of simulations for a fairer comparison.
start_time <- Sys.time()
res <- simulate_LD(1)
end_time <- Sys.time()
100*(end_time - start_time)
```

```
## Time difference of 2.281094 secs
```

```
apply(res, c(1, 3), mean, na.rm = TRUE)
```

```
##             estimate        2.5%       97.5%
## Intercept -117.252333 -70.1050789 -22.9578246
## Wind        -4.339366  -3.0516077  -1.7638492
## Temp         1.572657   2.0984399   2.6242233
## Month       -6.479740  -3.5209035  -0.5620666
## Day         -0.180512   0.2746808   0.7298737
```