# 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_MI2 <- function(runs = 100) {
  airquality_processed <- airquality %>% select(Ozone, Wind, Temp, Month, Day)
  res <- array(NA, dim = c(5, runs, 3))
  times <- array(NA, dim = c(100, 1, 1))
  dimnames(res) <- list(c("Intercept", "Wind", "Temp", "Month", "Day"),
                        as.character(1:runs), c("estimate", "2.5%", "97.5%"))
  for (run in 1:runs){
      # Note that time is only measured for the MI/imp steps
      # (i.e. filtering, predicting)
    start_time <- Sys.time()
    imp_MI <- mice(airquality_processed, print = FALSE)
    fit <- with(imp_MI, lm(Ozone ~ Wind + Temp + Month + Day))
    end_time <- Sys.time()
    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 %")])

    times[run, 1, 1] <- as.numeric(end_time - start_time)
  }
  list(res, times)
}
```

```
# Run 100 iterations of multiple imputations and store
res_MI2 <- simulate_MI2(100)
```

```
# Obtain confidence intervals & estimates for all coefficients, intercept.
apply(res_MI2[[1]], c(1, 3), mean, na.rm = TRUE)
```

```
##               estimate         2.5%       97.5%
## Intercept -61.4307377 -109.1171872 -13.7442882
## Wind       -3.1023161   -4.4656005  -1.7390316
## Temp        2.0040766    1.4704654   2.5376879
## Month      -3.6164996   -6.6794577  -0.5535415
## Day         0.2441269   -0.2167815   0.7050353
```

```r
# Mean time for iterations of multiple imputation
times <- res_MI2[[2]]
mean(times)
```

```
## [1] 0.07248978
```

### Listwise Deletion

```r
simulate_LD <- function(runs = 100){
  res <- array(NA, dim = c(5, 1, 3))
  dimnames(res) <- list(c("Intercept", "Wind", "Temp", "Month", "Day"),
                        as.character(1), c("estimate", "2.5%", "97.5%"))
  times <- array(NA, dim = c(runs, 1, 1))
  # Note that time is only measured for the LD/imp steps (i.e. filtering, predicting)
  for (run in 1:runs){
    start_time <- Sys.time()
    lw_airquality <- airquality %>% select(Ozone, Wind, Temp, Month, Day) %>%
      filter(!is.na(Ozone))
    fit <- with(lw_airquality, lm(Ozone ~ Wind + Temp + Month + Day))
    end_time <- Sys.time()
    times[run, 1, 1] <- as.numeric(end_time - start_time)
    # loop over each variable. Note we do the imputation just ONCE b/c LD is
    # deterministic.
    if (run == 1){
      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, 1, ] <- interval
      }
    }

  }
  list(res, times)
}
```

```r
result_LD <- simulate_LD()
```

```r
# Obtain confidence intervals & estimates for all coefficients, intercept.
apply(result_LD[[1]], 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
```

```r
# Mean time for 100 instances of LD
times_LD <- result_LD[[2]]
mean(times_LD)
```

```
## [1] 0.006822221
```