Missing Data

Brenton Kenkel — PSCI 8357
April 21, 2016

Varieties of Missingness

```
mc_listwise <- foreach(i = 1:1000, .combine = "rbind") %do% {
    dat <- sim_data(100) %>%
        mutate(x2 = ifelse(missing == 1, NA, x2))
    cf <- do(dat %>% group_by(type),
             lm(y ~ x1 + x2, data = .) %>% tidy())
    cf
}

mc_listwise %>%
    group_by(type, term) %>%
    summarise(mean = mean(estimate),
              se_true = sd(estimate),
              se_nominal = mean(std.error)) %>%
    mutate_each(funs(round(., 2)), -type, -term)
```

```
## Source: local data frame [12 x 5]
## Groups: type
##
##        type        term mean se_true se_nominal
## 1  Complete (Intercept) 1.01    0.10       0.10
## 2  Complete          x1 1.01    0.22       0.23
## 3  Complete          x2 0.99    0.20       0.20
## 4      MCAR (Intercept) 1.01    0.15       0.14
## 5      MCAR          x1 1.02    0.33       0.33
## 6      MCAR          x2 0.99    0.29       0.29
## 7       MAR (Intercept) 1.01    0.18       0.18
## 8       MAR          x1 1.01    0.33       0.35
## 9       MAR          x2 0.99    0.29       0.30
## 10       NI (Intercept) 1.00    0.19       0.18
```

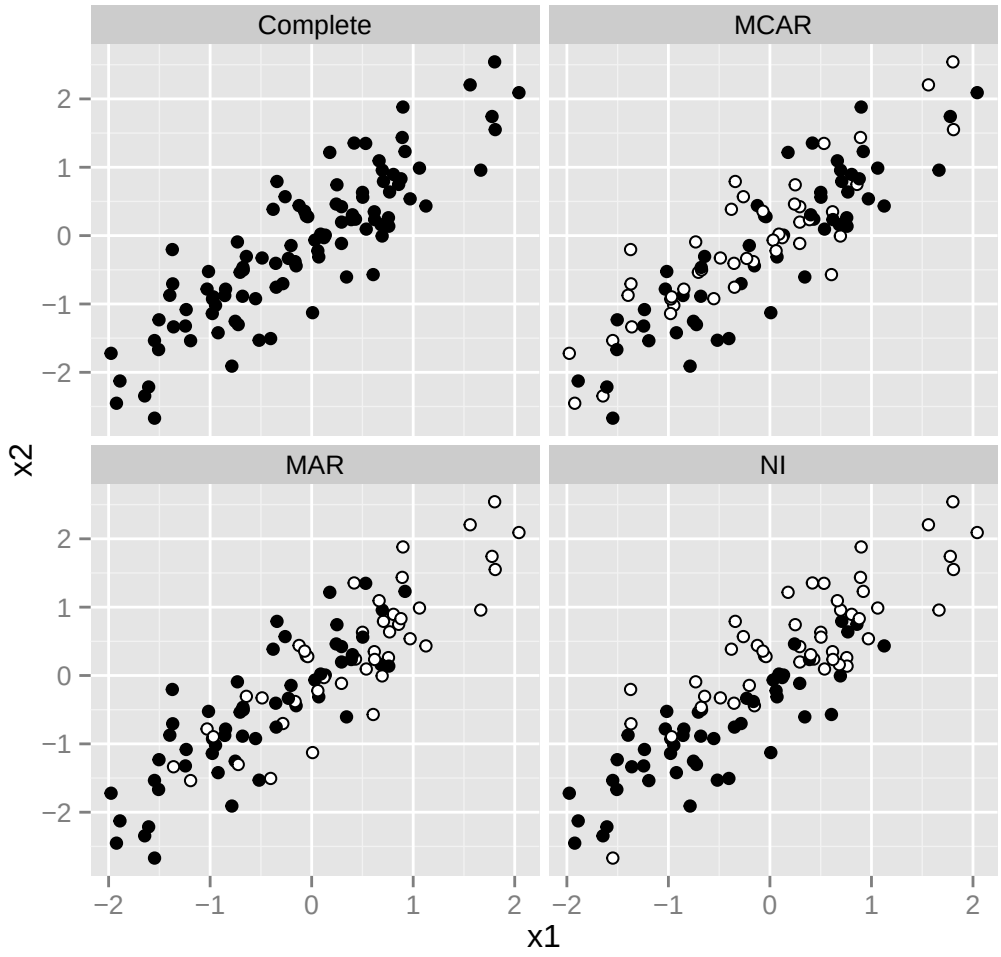

Figure 1. Illustration of data that are complete, missing completely at random (MCAR), missing at random (MAR), and nonignorably missing (NI).

```
## 11       NI          x1 1.01    0.33       0.33
## 12       NI          x2 0.99    0.32       0.31
```

```
mc_single <- foreach(i = 1:1000, .combine = "rbind") %do% {
    dat <- sim_data(100) %>%
        mutate(x2 = ifelse(missing == 1, NA, x2))

    cf <- do(dat %>% group_by(type), {
        x2_fit <- lm(x2 ~ x1, data = .)
        x2_pred <- predict(x2_fit, newdata = .)
        .$x2 <- ifelse(is.na(.$x2), x2_pred, .$x2)
        y_fit <- lm(y ~ x1 + x2, data = .)
        tidy(y_fit)
    })
}

mc_single %>%
    group_by(type, term) %>%
    summarise(mean = mean(estimate),
              se_true = sd(estimate),
              se_nominal = mean(std.error)) %>%
    mutate_each(funs(round(., 2)), -type, -term)
```

```
## Source: local data frame [12 x 5]
## Groups: type
##
##        type        term mean se_true se_nominal
## 1  Complete (Intercept) 1.00    0.10       0.10
## 2  Complete          x1 1.00    0.23       0.23
## 3  Complete          x2 1.00    0.20       0.20
## 4      MCAR (Intercept) 1.00    0.11       0.11
## 5      MCAR          x1 1.01    0.32       0.33
## 6      MCAR          x2 0.98    0.30       0.31
## 7       MAR (Intercept) 1.00    0.13       0.11
## 8       MAR          x1 0.99    0.33       0.33
## 9       MAR          x2 1.01    0.30       0.31
## 10       NI (Intercept) 1.18    0.14       0.12
## 11       NI          x1 1.09    0.31       0.32
## 12       NI          x2 1.01    0.31       0.33
```

References