Assignment 9b

Satvik Saha

2024-11-12

Answer 1

(a) The average effect in the population will be

$$0.35 \cdot 2.5 + 0.35 \cdot (-0.9) + 0.30 \cdot (1.9) = 1.13.$$

The standard error is given by

$$\sqrt{0.35^2 \cdot 1.1^2 + 0.35^2 \cdot 1.5^2 + 0.30^2 \cdot 1.2^2} = 0.74$$

(b) We solve

$$X = \frac{2.0 - 0.35 \cdot 3.5 - 0.35 \cdot (-1.9)}{0.30} = 4.8$$

Next,

$$Y = \sqrt{\frac{3.5^2 - 0.35^2 \cdot 2.1^2 - 0.35^2 \cdot 2.5^2}{0.30^2}} = 11.03.$$

Answer 2

(a) A complete-case analysis would only include the data points 8, 9, 10.

(b) For `y ~ z`, an available-case analysis would include the data points 5, 8, 9, 10.

For `y ~ x + z`, an available-case analysis would include the data points 8, 9, 10.

Answer 3

The estimated effect in future studies may be lower than in the initial study because of the following reasons.

1. *Change in conditions*: The effect may have changed (decreased) over time.
2. *Selection bias*: The initial study may have been published because a sufficiently large effect was observed. Future studies may not observe such a large effect if the true effect is lower.
3. *Dilution*: The initial study may have been focused on a subset of the population where the true effect is relatively large. Future studies which look at a larger population (or different parts of the population) may observe a lower effect.
4. *Incorrectness*: It is of course possible that the initial study was performed incorrectly in some way; perhaps the model was inappropriate or overfitted, or insufficient data were used. Future studies remedying these issues may observe something different.
5. *Malice*: Again, it is possible that the initial study was manipulated in some way, so as to obtain a large effect size which is not reflected in reality.

Research homework assignment

```r
df <- data.frame(
    x = c(NA, NA, NA, NA, NA, 1.11, 0.45, -1.55, 0.14, 0.11),
    y = c(NA, -0.60, NA, NA, 0.30, NA, NA, 1.40, -0.38, -0.82),
    z = c(1, NA, NA, 1, 0, 0, 2, 1, 1, 1)
)
df
```

```
##        x     y  z
## 1     NA    NA  1
## 2     NA -0.60 NA
## 3     NA    NA NA
## 4     NA    NA  1
## 5     NA  0.30  0
## 6   1.11    NA  0
## 7   0.45    NA  2
## 8  -1.55  1.40  1
## 9   0.14 -0.38  1
## 10  0.11 -0.82  1
```

```r
library(rstanarm)

impute <- function(df, maxiter = 1000, epsilon = 1e-3) {
    df.original <- df
    df.x.na <- is.na(df$x)
    df.y.na <- is.na(df$y)
    df.z.na <- is.na(df$z)

    df$x[df.x.na] <- rnorm(
        sum(df.x.na),
        mean = mean(df$x, na.rm = TRUE),
        sd = sd(df$x, na.rm = TRUE)
    )
    df$y[df.y.na] <- rnorm(
        sum(df.y.na),
        mean = mean(df$y, na.rm = TRUE),
        sd = sd(df$y, na.rm = TRUE)
    )
    df$z[df.z.na] <- rnorm(
        sum(df.z.na),
        mean = mean(df$z, na.rm = TRUE),
        sd = sd(df$z, na.rm = TRUE)
    )

    for (i in 1:maxiter) {
        x.old <- df$x[df.x.na]
        y.old <- df$y[df.y.na]
        z.old <- df$z[df.z.na]

        fit.x <- stan_glm(x ~ y + z, data = df[!df.x.na, ], refresh = FALSE)
        df$x[df.x.na] <- posterior_predict(fit.x, df[df.x.na, ], draws = 1)
        fit.y <- stan_glm(y ~ z + x, data = df[!df.y.na, ], refresh = FALSE)
        df$y[df.y.na] <- posterior_predict(fit.y, df[df.y.na, ], draws = 1)
        fit.z <- stan_glm(z ~ x + y, data = df[!df.z.na, ], refresh = FALSE)
        df$z[df.z.na] <- posterior_predict(fit.z, df[df.z.na, ], draws = 1)
        x.new <- df$x[df.x.na]
        y.new <- df$y[df.y.na]
        z.new <- df$z[df.z.na]

        x.score <- sd(x.new - x.old) / sd(df$x[!df.x.na])
        y.score <- sd(y.new - y.old) / sd(df$y[!df.y.na])
        z.score <- sd(z.new - z.old) / sd(df$z[!df.z.na])

        # print(c(x.score, y.score, z.score))
        if (max(c(x.score, y.score, z.score)) < epsilon) break
    }

    return(df)

}

# impute(df, maxiter = 20)
```

Issues: Convergence fails (at least, given a tractable number of iterations)!

Remark: Since this is in effect a Gibbs sampling procedure, the right notion of convergence might be in terms of the posterior distributions: for each row to be imputed, check whether the KL divergence between the new and old posterior distributions is sufficiently small. We shouldn't expect the posterior draws by themselves to be close!