that $X = W + \eta$ where $\eta \sim N(0, \tau^2 = 57)$. The goal is to fit the model $Y = \beta_o + \beta_1 X + \epsilon$ where $\epsilon \sim N(0, \sigma^2)$. Apply the multiple imputation methodology to construct point and interval estimates of β_1.

Site	Yield (Y)	Soil Nitrogen (W)	Site	Yield (Y)	Soil Nitrogen (W)
1	86	70	7	99	50
2	115	97	8	96	70
3	90	53	9	99	94
4	86	64	10	104	69
5	110	95	11	96	51
6	91	64			

Note that α_o, α_1 and τ^2 are known in the set of equations (8.2).

2. **Project:** Generate a sample of size 1000, $(X_i, Y_i, W_i), i = 1, 2, \ldots 1000$ from a trivariate normal distribution mean $(0, 1, 2)$ and covariance matrix,

$$\begin{bmatrix} 1 & 1 & 1.7 \\ & 2 & 1.5 \\ & & 4 \end{bmatrix}$$

   (a) The goal is to infer about the regression coefficient for X in the model $Y = \beta_0 + \beta_1 X + \epsilon$, $\epsilon \sim N(0, \sigma^2)$. Set aside the first 900 observations as data from the main study and treat the remaining 100 observations as data from a substudy. Fit the above regression model on the main study data and store the point estimates of β_o, β_1 and σ and the interval estimate of β_1.

   (b) Create a data corresponding to scenario (a) in Figure 8.1 by deleting X from the main study. Multiply impute the missing values of X in the main study. Perform multiply imputed analysis and again store the point and interval estimates of the same parameters.

   (c) Create a data corresponding to scenario (b) by deleting X values from the main study and Y values from the substudy. Multiply impute the missing values of X in the main study. Perform multiply imputed analysis as in (b).

   (d) Create a data corresponding to scenario (c) by deleting X on the main study, Y from the first 50 subjects in the substudy and W from the last 50 subjects. Perform multiply imputed analysis as in (b).

(e) Generate new samples and repeat the process (a) to (d), 250 times.

(f) Compare the bias and mean square properties of the estimates of β_o, β_1 and σ^2.

(g) Compute the true value of β_1 and calculate the actual coverage rate for each method of estimating the confidence interval. Also, calculate the length of the confidence intervals.

Based on this simulation study write a brief report on your findings and recommendations.

3. Refer to the example in Section 8.3.1. Suppose that one is interested in assessing the effect of not raking the generated values Z for the nonsampled subjects to the known population total. Construct inference about Q, by redoing the analysis of data without raking.

4. Sukhatme and Sukhatme (1970) provide area under wheat for 1936, X, as well. Construct inference about Q by using both X and Z as covariates.

5. **Project:** From the NHANES series, identify six years and a set of four variables (X_1, X_2, X_3, X_4) that were collected every year. Call this a complete data. From year 1, delete all values except for variables X_1 and X_2. Similarly, keep (X_1, X_3) in year 2; in year 3, (X_1, X_4)); in year 4 (X_2, X_3); in year 5, (X_2, X_4); and, finally, in year 6, keep (X_3, X_4). Vertically concatenate the six data sets. Multiply impute the missing values in the concatenated data set. Perform several example analysis that require 3 or more variables from the multiply imputed data sets and also on vertically concatenated six complete data sets. Compare the point and interval estimates. Write a brief report on your findings.

6. This problem is adapted based on a study described in Fleiss (1986). A large nursing home has a population of about 400 patients with senile dementia. Two methods (A and B) for training patients to take care of themselves were under consideration. A randomized study was conducted with 11 patients receiving training method A and 8 receiving training method B. Two weeks after training, each