**Simulation Exercises**

Day 1A

1. Generate 1000 iid realizations of (Y, X, V) from a known joint distribution:

V ~ N(0,1); X|V ~ N(V,1); Y|X,V ~ N(X – V, 1)

1. Estimate the following:
   1. Regression coefficients for the model E(Y|X,V)= + X + V
   2. Regression coefficients for the model E(Y|X)= + X
   3. The mean of X

Estimate ( ), , and E(X) among those without missing data in the following scenarios:

1. Create MCAR data in X with approximately 50% missing (i.e., P(R|X, Y, V) = P(R)).
2. Create ~50% missing data in X such that P(R|X, Y, V) = P(R|V).
3. Create ~50% missing data in X such that P(R|X, Y, V) = P(R|X).
4. Create ~50% missing data in X such that P(R|X, Y, V) = P(R|Y).

Compare estimates in each setting with those under the full data model (question 2). Are they close? How do their standard errors compare?

Day 1B

Using the 4 simulated datasets you generated earlier today, obtain regression estimates for the model E(Y|X,V)= + X + V using the following methods:

1. Mean imputation
2. Conditional expectation imputation
3. Single imputation from the fitted distribution
4. Missing indicator approach
5. Bonus: Hot deck imputation

Compare and contrast estimates with estimates based on the original data based on the complete case analysis.

Day 2A

Fit an inverse probability weighted (IPW) estimator of the mean of X for the following data you generated on Day 1:

1. Missing data that are MCAR
2. Missing data where Rx (X, Y) | V
3. Missing data where Rx (V, Y) | X
4. Missing data where Rx (V, X) | Y

How do these estimates compare to the mean of X based on the full data?

Day 2B

Fit an IPW estimator for the regression coefficients of the model

E(Y|X,V)= + X + V

using the data you generated on Day 1.

1. Missing data that are MCAR
2. Missing data where Rx (X, Y) | V
3. Missing data where Rx (V, Y) | X
4. Missing data where Rx (V, X) | Y

How do these estimates compare to the regression estimates based on the full data?

Day 2C

Compute standard errors and 95% confidence intervals using the bootstrap for the IPW regression estimates of the model

E(Y|X,V)= + X + V

using the data you generated on Day 1:

1. Missing data that are MCAR
2. Missing data where Rx (X, Y) | V
3. Missing data where Rx (V, Y) | X
4. Missing data where Rx (V, X) | Y

How do these standard errors and 95% confidence intervals compare to those that do not use the bootstrap?

Day 3A

For the simulated data in Day 1 with MCAR, code the MI with the “Predict” method for the missing using model:

Remember three steps (impute, analysis, pool).

1. How do these regression coefficients and their standard errors compare with those using the full data?

Day 3B

For the simulated data in Day 3A, code the MI with the "Predict+noise" method for the missing using model:

1. Compare these regression coefficients and their standard errors with those using the full data and the “Predict” method.

Day 3C

For the simulated data in Day 3A, code the MI with the "predict + noise + parameters uncertainty" method for the missing using model:

where , , and , , is the variance estimate from the observed data, and is its degrees of freedom.

1. Compare these regression coefficients and their standard errors with those using the full data and results from 3A and 3B.
2. How do they compare with the IPW estimators that you obtained on Day 2?
3. How do these estimates change if you do not include Y in your imputation model?

Day 3D

Fit a multiple imputation estimator using the mice package with Bayesian MI approach in R assuming normality for the data generated in Day 3A.

1. How do these regression coefficients and their standard errors compare with those that you calculated by hand using Bayesian MI approach (3C)?

Day 3E:

For the data generated in Exercise Day 3A, use PMM method and compare the results from Day 3D.

Day 4A

For exercise Day 3C, part 3, now remove outcome y from the imputation model for x using the mice package.

Day 4B

Generate 1000 iid realizations of (Y, X, V) from a known joint distribution:

V ~ Bernoulli(0.5); X|V ~ N(V,1); Y|X,V ~ N(X – V, 1)

Generate Rv ~ Bernoulli(0.5) and Rx ~ Bernoulli(px) where px = expit(V)

Estimate regression coefficients for the model E(Y|X,V)= + X + V in the following manner:

1. Complete case analysis
2. Multiple imputation using default settings in mice
3. Multiple imputation using other settings in mice

How do the coefficient estimates and their standard errors differ between the three approaches?