Revisiting the Available Cases Method for Missing Values

Xiao (Max) Gu and Norm Matloff
University of California at Davis

JSM 2015

Xiao (Max) Gu and Norm Matloff
University of California at Davis

Taxonomy of Methods

Taxonomy of Methods

Major current methods:

Xiao (Max) Gu and Norm Matloff
University of California at Davis

Taxonomy of Methods

Major current methods:

- Use only complete cases (CC).
- Multiple imputation (MI).
- MLE.

Taxonomy of Methods

Major current methods:

- Use only complete cases (CC).
- Multiple imputation (MI).
- MLE.

Forgotten method:

Taxonomy of Methods

Major current methods:

- Use only complete cases (CC).
- Multiple imputation (MI).
- MLE.

Forgotten method:

- Available cases (AC). Use partially-intact cases when possible.

Xiao (Max) Gu and Norm Matloff
University of California at Davis

Overview of AC Method

Overview of AC Method

E.g. linear regreesion (random-X).

Overview of AC Method

E.g. linear regreesion (random-X).

$$\widehat{\beta} = (X'X)^{-1}X'Y = \left[\frac{1}{n}(X'X)^{-1}\right]\left[\frac{1}{n}X'Y\right] = A^{-1}D \qquad (1)$$

Xiao (Max) Gu and Norm Matloff University of California at Davis

Overview of AC Method

E.g. linear regreesion (random-X).

$$\widehat{\beta} = (X'X)^{-1}X'Y = \left[\frac{1}{n}(X'X)^{-1}\right]\left[\frac{1}{n}X'Y\right] = A^{-1}D \qquad (1)$$

A estimates quantities like

Xiao (Max) Gu and Norm Matloff
University of California at Davis

Overview of AC Method

E.g. linear regreesion (random-X).

$$\widehat{\beta} = (X'X)^{-1}X'Y = \left[\frac{1}{n}(X'X)^{-1}\right]\left[\frac{1}{n}X'Y\right] = A^{-1}D \qquad (1)$$

A estimates quantities like

$$E[X^{(i)}X^{(j)}] \qquad (2)$$

Overview of AC Method

E.g. linear regreesion (random-X).

$$\widehat{\beta} = (X'X)^{-1}X'Y = \left[\frac{1}{n}(X'X)^{-1}\right]\left[\frac{1}{n}X'Y\right] = A^{-1}D \tag{1}$$

A estimates quantities like

$$E[X^{(i)}X^{(j)}] \tag{2}$$

while D estimates quanatities like

Overview of AC Method

E.g. linear regreesion (random-X).

$$\widehat{\beta} = (X'X)^{-1}X'Y = \left[\frac{1}{n}(X'X)^{-1}\right]\left[\frac{1}{n}X'Y\right] = A^{-1}D \quad (1)$$

A estimates quantities like

$$E[X^{(i)}X^{(j)}] \quad (2)$$

while D estimates quanatities like

$$E[X^{(i)}Y] \quad (3)$$

Xiao (Max) Gu and Norm Matloff
University of California at Davis

AC Overview, cont'd.

Xiao (Max) Gu and Norm Matloff
University of California at Davis

AC Overview, cont'd.

CC seems wasteful.

AC Overview, cont'd.

CC seems wasteful.

- In estimating, say, $E[X^{(2)}Y]$, why throw out cases in which $X^{(2)}$ and Y are intact but $X^{(5)}$ is missing?

AC Overview, cont'd.

CC seems wasteful.

- In estimating, say, $E[X^{(2)}Y]$, why throw out cases in which $X^{(2)}$ and Y are intact but $X^{(5)}$ is missing?
- Instead, estimate by $E[X^{(i)}Y]$ by

$$\frac{1}{M} \sum_{X^{(i)},\ Y \text{ intact}} X_k^{(i)} Y_k \tag{4}$$

where $M = \#$ of cases with both $X^{(i)}$ and Y intact.

AC Overview, cont'd.

CC seems wasteful.

- In estimating, say, $E[X^{(2)}Y]$, why throw out cases in which $X^{(2)}$ and Y are intact but $X^{(5)}$ is missing?
- Instead, estimate by $E[X^{(i)}Y]$ by

$$\frac{1}{M} \sum_{X^{(i)},\ Y \text{ intact}} X_k^{(i)} Y_k \tag{4}$$

where $M = \#$ of cases with both $X^{(i)}$ and Y intact.
- Same for the quantities $E[X^{(i)}X^{(j)}]$.

Xiao (Max) Gu and Norm Matloff
University of California at Davis

AC Sounds Good, But Not Popular

AC Sounds Good, But Not Popular

- AC should be more accurate that CC — uses more data.

AC Sounds Good, But Not Popular

- AC should be more accurate that CC — uses more data.
- Yet, AC seems to have been dismissed early on in the Missing Value literature,

AC Sounds Good, But Not Popular

- AC should be more accurate that CC — uses more data.
- Yet, AC seems to have been dismissed early on in the Missing Value literature, apparently because:

AC Sounds Good, But Not Popular

- AC should be more accurate that CC — uses more data.
- Yet, AC seems to have been dismissed early on in the Missing Value literature, apparently because:
  - The modified $X'X$ may not be positive definite.

AC Sounds Good, But Not Popular

- AC should be more accurate that CC — uses more data.
- Yet, AC seems to have been dismissed early on in the Missing Value literature, apparently because:
  - The modified $X'X$ may not be positive definite.
  - AC assumes MCAR, the strongest among the famous assumption sets.

AC Sounds Good, But Not Popular

- AC should be more accurate that CC — uses more data.
- Yet, AC seems to have been dismissed early on in the Missing Value literature, apparently because:
  - The modified $X'X$ may not be positive definite.
  - AC assumes MCAR, the strongest among the famous assumption sets.
- Still, AC seems worth revisiting.

AC Sounds Good, But Not Popular

- AC should be more accurate that CC — uses more data.
- Yet, AC seems to have been dismissed early on in the Missing Value literature, apparently because:
  - The modified $X'X$ may not be positive definite.
  - AC assumes MCAR, the strongest among the famous assumption sets.
- Still, AC seems worth revisiting.
  - Lack of positive definiteness is unlikely to occur, and it's unclear whether it's important anyway.

AC Sounds Good, But Not Popular

- AC should be more accurate that CC — uses more data.
- Yet, AC seems to have been dismissed early on in the Missing Value literature, apparently because:
  - The modified $X'X$ may not be positive definite.
  - AC assumes MCAR, the strongest among the famous assumption sets.
- Still, AC seems worth revisiting.
  - Lack of positive definiteness is unlikely to occur, and it's unclear whether it's important anyway.
  - The most common alternative assumption set, MAR, is also quite strong.

AC Sounds Good, But Not Popular

- AC should be more accurate that CC — uses more data.
- Yet, AC seems to have been dismissed early on in the Missing Value literature, apparently because:
  - The modified $X'X$ may not be positive definite.
  - AC assumes MCAR, the strongest among the famous assumption sets.
- Still, AC seems worth revisiting.
  - Lack of positive definiteness is unlikely to occur, and it's unclear whether it's important anyway.
  - The most common alternative assumption set, MAR, is also quite strong. (More on this later.)

Our Study: AC vs. CC, MI

Our Study: AC vs. CC, MI

- Here we “reopen the case” regarding AC,

Our Study: AC vs. CC, MI

- Here we “reopen the case” regarding AC, comparing to CC and MI.

Our Study: AC vs. CC, MI

- Here we “reopen the case” regarding AC, comparing to CC and MI.
- We look at the old application, linear regression,

Our Study: AC vs. CC, MI

- Here we "reopen the case" regarding AC, comparing to CC and MI.
- We look at the old application, linear regression, and 2 new ones: PCA and log-linear model.

Our Study: AC vs. CC, MI

- Here we "reopen the case" regarding AC, comparing to CC and MI.
- We look at the old application, linear regression, and 2 new ones: PCA and log-linear model.
- We look at these criteria:

Our Study: AC vs. CC, MI

- Here we "reopen the case" regarding AC, comparing to CC and MI.
- We look at the old application, linear regression, and 2 new ones: PCA and log-linear model.
- We look at these criteria:
  - Applicability.

Our Study: AC vs. CC, MI

- Here we "reopen the case" regarding AC, comparing to CC and MI.
- We look at the old application, linear regression, and 2 new ones: PCA and log-linear model.
- We look at these criteria:
    - Applicability.
    - Variance, bias.

Our Study: AC vs. CC, MI

- Here we "reopen the case" regarding AC, comparing to CC and MI.
- We look at the old application, linear regression, and 2 new ones: PCA and log-linear model.
- We look at these criteria:
    - Applicability.
    - Variance, bias.
    - Run time.

Our Study: AC vs. CC, MI

- Here we "reopen the case" regarding AC, comparing to CC and MI.
- We look at the old application, linear regression, and 2 new ones: PCA and log-linear model.
- We look at these criteria:
  - Applicability.
  - Variance, bias.
  - Run time.
- For MI, we use Amelia 2.

Xiao (Max) Gu and Norm Matloff
University of California at Davis

Linear Regression

Linear Regression

- All 3 methods are applicable.

Linear Regression

- All 3 methods are applicable.
- Simulation results: n = 10000, p = 3, 10% missing, $\beta_1 = 1$

Linear Regression

- All 3 methods are applicable.
- Simulation results: n = 10000, p = 3, 10% missing, $\beta_1 = 1$

method	mean	variance	time
CC	0.9996	0.0002	0.79
MI	0.9784	0.0002	142.02
AC	1.0027	0.0010	23.80

Note: Most time in AC spent in finding numeric derivs for standard errors.

Linear Regression

- All 3 methods are applicable.
- Simulation results: n = 10000, p = 3, 10% missing, $\beta_1 = 1$

method	mean	variance	time
CC	0.9996	0.0002	0.79
MI	0.9784	0.0002	142.02
AC	1.0027	0.0010	23.80

  Note: Most time in AC spent in finding numeric derivs for standard errors.
- MI slightly biased.

Linear Regression

- All 3 methods are applicable.
- Simulation results: n = 10000, p = 3, 10% missing, $\beta_1 = 1$

method	mean	variance	time
CC	0.9996	0.0002	0.79
MI	0.9784	0.0002	142.02
AC	1.0027	0.0010	23.80

  Note: Most time in AC spent in finding numeric derivs for standard errors.
- MI slightly biased.
- AC terrible MSE. (Some intuition....)

Linear Regression

- All 3 methods are applicable.
- Simulation results: n = 10000, p = 3, 10% missing, $\beta_1 = 1$

method	mean	variance	time
CC	0.9996	0.0002	0.79
MI	0.9784	0.0002	142.02
AC	1.0027	0.0010	23.80

  Note: Most time in AC spent in finding numeric derivs for standard errors.
- MI slightly biased.
- AC terrible MSE. (Some intuition....)
- MI terrible run time.

Linear Regression

- All 3 methods are applicable.
- Simulation results: n = 10000, p = 3, 10% missing, $\beta_1 = 1$

method	mean	variance	time
CC	0.9996	0.0002	0.79
MI	0.9784	0.0002	142.02
AC	1.0027	0.0010	23.80

  Note: Most time in AC spent in finding numeric derivs for standard errors.
- MI slightly biased.
- AC terrible MSE. (Some intuition....)
- MI terrible run time.
- Verdict: Use CC.

PCA

Xiao (Max) Gu and Norm Matloff
University of California at Davis

PCA

- CC, AC methods applicable.

PCA

- CC, AC methods applicable.
- MI sometimes gave error message ("perfectly collinear...").

PCA

- CC, AC methods applicable.
- MI sometimes gave error message ("perfectly collinear...").

  .
- Simulation results: n = 100, p = 10, 10% missing; largest eigenvalue; ρ matrix

PCA

- CC, AC methods applicable.
- MI sometimes gave error message ("perfectly collinear...").

.

- Simulation results: n = 100, p = 10, 10% missing; largest eigenvalue; ρ matrix

method	mean	variance
CC	2.3328	0.0517
AC	2.1012	0.0218

PCA

- CC, AC methods applicable.
- MI sometimes gave error message ("perfectly collinear...").

  .
- Simulation results: $n = 100$, $p = 10$, 10% missing; largest eigenvalue; ρ matrix

method	mean	variance
CC	2.3328	0.0517
AC	2.1012	0.0218

Xiao (Max) Gu and Norm Matloff
University of California at Davis

A Note on PCA

A Note on PCA

- PCA is upward biased anyway (even with no NAs), since PCA naturally overfits.

Xiao (Max) Gu and Norm Matloff
University of California at Davis

A Note on PCA

- PCA is upward biased anyway (even with no NAs), since PCA naturally overfits. (First comp. maxes var. of lin. combs. of length 1.)

A Note on PCA

- PCA is upward biased anyway (even with no NAs), since PCA naturally overfits. (First comp. maxes var. of lin. combs. of length 1.)
- The means of 2.1 and 2.3 we got for $n = 100$ become about 1.97 for $n = 1000$.

A Note on PCA

- PCA is upward biased anyway (even with no NAs), since PCA naturally overfits. (First comp. maxes var. of lin. combs. of length 1.)
- The means of 2.1 and 2.3 we got for $n = 100$ become about 1.97 for $n = 1000$.
- But in all simulation runs, AC was *less* upward biased, and had small variance, compared to CC.

A Note on PCA

- PCA is upward biased anyway (even with no NAs), since PCA naturally overfits. (First comp. maxes var. of lin. combs. of length 1.)
- The means of 2.1 and 2.3 we got for $n = 100$ become about 1.97 for $n = 1000$.
- But in all simulation runs, AC was *less* upward biased, and had small variance, compared to CC. This was severe for larger values of p.

Contingency Table Models

Contingency Table Models

- MI not appropriate, since assumes MV normal data.

Contingency Table Models

- MI not appropriate, since assumes MV normal data. (Though MI methods do exist for this setting.)

Contingency Table Models

- MI not appropriate, since assumes MV normal data. (Though MI methods do exist for this setting.)
- Example: Factors X, Y, Z;

Contingency Table Models

- MI not appropriate, since assumes MV normal data. (Though MI methods do exist for this setting.)
- Example: Factors X, Y, Z; (12)(13) model — Y and Z independent, given X.

Contingency Table Models

- MI not appropriate, since assumes MV normal data. (Though MI methods do exist for this setting.)
- Example: Factors X, Y, Z; (12)(13) model — Y and Z independent, given X.
- In terms of marginal distributions:

$$p_{ijk} = p_{i..} \frac{p_{i.j}}{p_{i..}} \frac{p_{i.k}}{p_{i..}} = \frac{p_{i.j} p_{i.k}}{p_{i..}} \qquad (5)$$

- E.g. set $\widehat{p}_{i.k}$ to the proportion of cases in which $X = i, \ Z = k$, among cases in which X and Z are intact.
- Simulation example: (1)(23) model, n = 100, est. p_{111}.

method	mean	var
CC	0.1246591	0.0009020450
AC	0.1249168	0.0007548656

Contingency Table Models

- MI not appropriate, since assumes MV normal data. (Though MI methods do exist for this setting.)
- Example: Factors X, Y, Z; (12)(13) model — Y and Z independent, given X.
- In terms of marginal distributions:

$$p_{ijk} = p_{i..} \frac{p_{i.j}}{p_{i..}} \frac{p_{i.k}}{p_{i..}} = \frac{p_{i.j} p_{i.k}}{p_{i..}} \tag{5}$$

- E.g. set $\widehat{p}_{i.k}$ to the proportion of cases in which $X = i,\ Z = k$, among cases in which X and Z are intact.
- Simulation example: (1)(23) model, n = 100, est. p_{111}.

method	mean	var
CC	0.1246591	0.0009020450
AC	0.1249168	0.0007548656

AC advantage more if have more factors or higher NA %.

Xiao (Max) Gu and Norm Matloff
University of California at Davis

On Assumptions

On Assumptions

- CC, AC assume MCAR, stronger than MI's MAR.

Xiao (Max) Gu and Norm Matloff
University of California at Davis

On Assumptions

- CC, AC assume MCAR, stronger than MI's MAR.
- However:

On Assumptions

- CC, AC assume MCAR, stronger than MI's MAR.
- However:
  - Arguably, $\mathrm{MAR} \cap \mathrm{MCAR}^c$ rare in practice.

On Assumptions

- CC, AC assume MCAR, stronger than MI's MAR.
- However:
    - Arguably, $\mathrm{MAR} \cap \mathrm{MCAR}^c$ rare in practice.
    - $\widehat{\beta}$ still unbiased for β under CC, AC even under $\mathrm{MAR} \cap \mathrm{MCAR}^c$.

On Assumptions

- CC, AC assume MCAR, stronger than MI's MAR.
- However:
  - Arguably, $\mathrm{MAR} \cap \mathrm{MCAR}^c$ rare in practice.
  - $\widehat{\beta}$ still unbiased for β under CC, AC even under $\mathrm{MAR} \cap \mathrm{MCAR}^c$.
  - In $\mathrm{MAR} \cap \mathrm{MCAR}^c$ case, bias does arise if use CC or AC to estimate EY or $EX^{(i)}$.

On Assumptions

- CC, AC assume MCAR, stronger than MI's MAR.
- However:
  - Arguably, $\mathrm{MAR} \cap \mathrm{MCAR}^c$ rare in practice.
  - $\widehat{\beta}$ still unbiased for β under CC, AC even under $\mathrm{MAR} \cap \mathrm{MCAR}^c$.
  - In $\mathrm{MAR} \cap \mathrm{MCAR}^c$ case, bias does arise if use CC or AC to estimate EY or $EX^{(i)}$. In such case, use Matloff, *Biometrika*, 1982.

Xiao (Max) Gu and Norm Matloff
University of California at Davis

Software

Software

- Code available at *https://github.com/maxguxiao/Available-Cases.git*. Currently under development; check current status.

Software

- Code available at *https://github.com/maxguxiao/Available-Cases.git*. Currently under development; check current status.
- R's **cov()**, **cor()** functions include the option **use = 'pairwise.complete.obs'**, which is the AC method.

Software

- Code available at *https://github.com/maxguxiao/Available-Cases.git*. Currently under development; check current status.
- R's **cov()**, **cor()** functions include the option **use = 'pairwise.complete.obs'**, which is the AC method. This could be used to implement AC in two applications:

Software

- Code available at *https://github.com/maxguxiao/Available-Cases.git*. Currently under development; check current status.
- R's **cov()**, **cor()** functions include the option **use = 'pairwise.complete.obs'**, which is the AC method. This could be used to implement AC in two applications:
    - For PCA, just run **eigen()** on either a covariance or correlation matrix computed for AC as above.

Software

- Code available at *https://github.com/maxguxiao/Available-Cases.git*. Currently under development; check current status.
- R's **cov()**, **cor()** functions include the option **use = 'pairwise.complete.obs'**, which is the AC method. This could be used to implement AC in two applications:
  - For PCA, just run **eigen()** on either a covariance or correlation matrix computed for AC as above.
  - For linear regression, the matrices A and D both can be computed using **cov()**, after adjusting via a centering operation.

Software

- Code available at *https://github.com/maxguxiao/Available-Cases.git*. Currently under development; check current status.
- R's **cov()**, **cor()** functions include the option **use = 'pairwise.complete.obs'**, which is the AC method. This could be used to implement AC in two applications:
    - For PCA, just run **eigen()** on either a covariance or correlation matrix computed for AC as above.
    - For linear regression, the matrices A and D both can be computed using **cov()**, after adjusting via a centering operation.

Conclusions

Conclusions

- Final score: AC had 2 wins, 1 loss.

Xiao (Max) Gu and Norm Matloff
University of California at Davis

Conclusions

- Final score: AC had 2 wins, 1 loss.
- MI quite time-consuming, not recommended unless MCAR an issue.

Conclusions

- Final score: AC had 2 wins, 1 loss.
- MI quite time-consuming, not recommended unless MCAR an issue.

These slides available at
http://heather.cs.ucdavis.edu/SeattleSlides.pdf