Strategies for Missing Data in Educational Research

Dan McNeish, Laura M. Stapleton, & Alison Preston
University of Maryland, College Park

Outline

- Introduction and context
- Traditional methods for handling missing data and their shortcomings
- Maximum Likelihood methods for missing data
- Multiple imputation methods for missing data

The missing data problem

- Missing values are a ubiquitous problem in nearly all data sets
- They can either be a source of large biases or can be completely innocuous
  - Cannot determine which because there are no values to inspect
- As a result, dealing with missing data is a notoriously difficult problem when analyzing data

Examples

- Imagine that some people's wage information is missing in a dataset that only contains employed individuals.

- It could be missing because :
  - A data entry error where some values did not make it into the dataset
  - The individual worked in a sector that is not required to report wage information
  - The individual was embarrassed to report a low wage

Examples

Each of these scenarios present a different type of missing data that each have their own unique challenges

- It could be missing because :
  - A data entry error where some values did not make it into the dataset
  - The individual worked in a sector that is not required to report wage information
  - The individual was embarrassed to report a low wages

Conventional ways to classify missingness

- Missing Completely at Random (MCAR)
  - This is the ideal case but rarely seen in practice
  - Usually the result of some problem with data collection or entry
- Missing at Random (MAR)
  - The missing value is related to some other variable that has been collected
- Missing Not at Random (MNAR)
  - The missing value is related to a variable that was not collected or not observed

Missing Completely at Random (MCAR)

- The observed data can be viewed as a random sub-sample of the overall data
- The probability that a value is missing is not affected by any observed or unobserved variables in the data
- MCAR missingness is the only type of missingness that can be explicitly tested

Numerical Example

Worker ID	IQ	Wage Decile
1	85	6
2	89	4
3	91	7
4	92	2
5	94	3
6	97	5
7	99	8
8	101	5
9	104	2
10	105	4
11	108	6
12	111	8
13	112	8
14	118	10
15	121	9

Perhaps two workers were sick and did not turn in their paperwork when data were collected so they have missing values

Numerical Example

Worker ID	IQ	Wage Decile
1	85	6
2	89	MISSING
3	91	7
4	92	2
5	94	3
6	97	5
7	99	8
8	101	5
9	104	2
10	105	4
11	108	6
12	111	MISSING
13	112	8
14	118	10
15	121	9

Perhaps two workers were sick and did not turn in their paperwork when data were collected so they have missing values

This missingness is MCAR – it does not depend on any other variables in the data, it occurred strictly by random chance

Missing at Random (MAR)

- When missing values are MAR, the missingness is related only to other variables in the data

- Somewhat confusing terminology since random seems to imply that the missingness is haphazard
  - Naming convention comes from probability, which is not always intuitive to non-probabilists

- There is no way to explicitly test for MAR, analysts must assume that missing values are due to other observed variables in the data

Worker ID	IQ	Wage Decile
1	85	6
2	89	4
3	91	7
4	92	2
5	94	3
6	97	5
7	99	8
8	101	5
9	104	2
10	105	4
11	108	6
12	111	8
13	112	8
14	118	10
15	121	9

Perhaps only workers who have IQs above 100 are able to find employment

Worker ID	IQ	Wage Decile
1	85	MISSING
2	89	MISSING
3	91	MISSING
4	92	MISSING
5	94	MISSING
6	97	MISSING
7	99	MISSING
8	101	5
9	104	2
10	105	4
11	108	6
12	111	8
13	112	8
14	118	10
15	121	9

Perhaps only workers who have IQs above 100 are able to find employment

The missing values are not haphazardly missing, but are missing in relation to another variable in the data (IQ)

MLDS CENTER
Maryland Longitudinal Data System
Better Data • Informed Choices • Improved Results

MISSING NOT AT RANDOM (MNAR)

- With MNAR missingness, the missing value for a variable depends on the value of the variable itself

- Best explained with examples

    - Not reporting age or weight because people may not want to reveal the information/are embarrassed

    - Skipping questions on risky behaviors in public health studies

Worker ID	IQ	Wage Decile
1	85	6
2	89	4
3	91	7
4	92	2
5	94	3
6	97	5
7	99	8
8	101	5
9	104	2
10	105	4
11	108	6
12	111	8
13	112	8
14	118	10
15	121	9

MLDS CENTER
Maryland Longitudinal Data System
Better Data • Informed Choices • Improved Results

People with the lowest wages chose not to report their earnings

Worker ID	IQ	Wage Decile
1	85	6
2	89	MISSING
3	91	7
4	92	MISSING
5	94	MISSING
6	97	5
7	99	8
8	101	5
9	104	MISSING
10	105	MISSING
11	108	6
12	111	8
13	112	8
14	118	10
15	121	9

People with the lowest wages chose not to report their earnings

The Wage Decile data is now missing based upon itself.

The value is missing because it would have been low had the data been reported.

Traditional Methods for Missing Data- Deletion

Listwise Deletion

- Deletes any observation that has a missing value for **any** variable in the analysis
  - Alternatively called complete-case analysis
- Listwise deletion is very convenient and has the advantage that a common set of observations is used for all analyses
- A common default for general software programs (e.g., SPSS, SAS)

Listwise Deletion II

- Deleting cases can be extremely wasteful, especially if there is even a moderate amount of missing data
  - If the data have 20 variables each with 2% missing, only about 67% of cases will be used
- Estimates will be unbiased with listwise deletion if missing data are MCAR
  - In this case, complete cases can be considered a random subset of the full data
  - With MNAR or MAR, listwise deletion gives biased estimates
- Although convenient, listwise deletion is only valid under very specific circumstances

Pairwise Deletion

- Similar to listwise deletion except that observations are only deleted if the variables directly involved in the analysis have missing data
- Idea is to conserve as many observations as possible by limiting the number of deleted cases
- This is most commonly seen with a correlations where each correlation is based on a different sample size

Listwise vs. Pairwise

ID	X1	X2	X3
1	10	25	---
2	11	24	---
3	15	29	34
4	12	24	25
5	---	21	40
6	9	---	41

- Suppose we want to calculate the correlation between X1 & X2, X2 & X3, and X1 & X3 for the data above

Listwise vs. Pairwise

ID	X1	X2	X3
~~1~~	~~10~~	~~25~~	~~---~~
~~2~~	~~11~~	~~24~~	~~---~~
3	15	29	34
4	12	24	25
~~5~~	~~---~~	~~21~~	~~40~~
~~6~~	~~9~~	~~---~~	~~41~~

- If we wanted the correlation between these variables, listwise deletion would remove IDs 1, 2, 5, and 6.
  - Listwise deletion removes any cases that are missing values on one or more variables

Listwise vs. Pairwise

ID	X1	X2
1	10	25
2	11	24
3	15	29
4	12	24
~~5~~	~~---~~	~~21~~
~~6~~	~~9~~	~~---~~

- Listwise – Corr(X1,X2) uses 2 observations
- Pairwise – Corr(X1,X2) uses 4 observations

Listwise vs. Pairwise

ID	X1	X3
~~1~~	~~10~~	~~---~~
~~2~~	~~11~~	~~---~~
3	15	34
4	12	25
~~5~~	~~---~~	~~40~~
6	9	41

- Listwise – Corr(X1,X3) uses 2 observations
- Pairwise – Corr(X1,X3) uses 3 observations

Listwise vs. Pairwise

ID	X2	X3
1	~~25~~	~~---~~
2	~~24~~	~~---~~
3	29	34
4	24	25
5	21	40
6	~~---~~	~~41~~

- Listwise – Corr(X2,X3) uses 2 observations
- Pairwise – Corr(X2,X3) uses 3 observations
- With listwise deletion, all cells have the same number of observations, all analyses based on same people
- Downside with pairwise is that every estimate is based on different people and each estimate has a different sample size

Traditional Methods for Missing Data- Single Imputation

General Idea

- Rather than deleting observations with missing cases, single imputation methods fill in the missing values with a statistical guess for what the value would have been

- Complete data are used in a variety of ways to come up with reasonable values for imputing the missing values
  - Mean Imputation
  - Regression Imputation

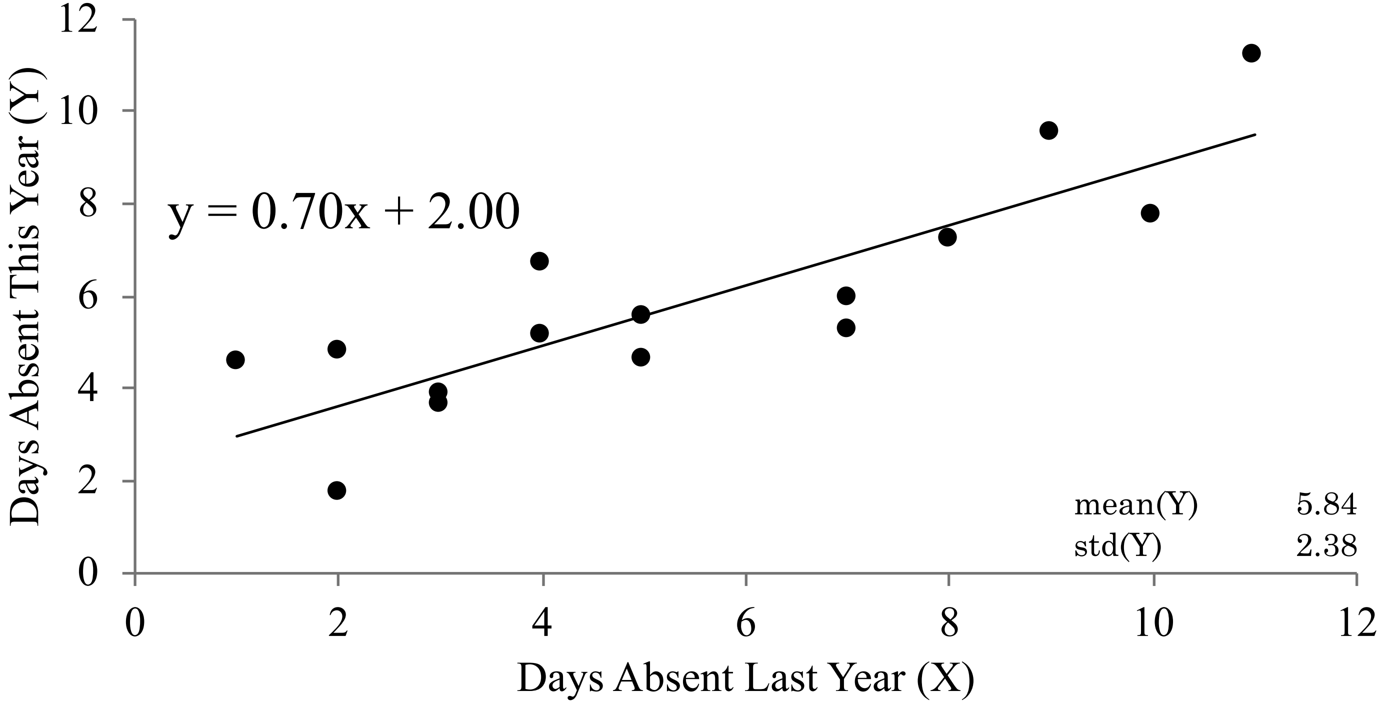
Mean Imputation

- Mean imputation replaces any missing values with the arithmetic mean across all individuals in the dataset for that variable

- This method is the worst possible choice, resulting estimates are almost always biased
  - Listwise deletion routinely outperforms mean imputation

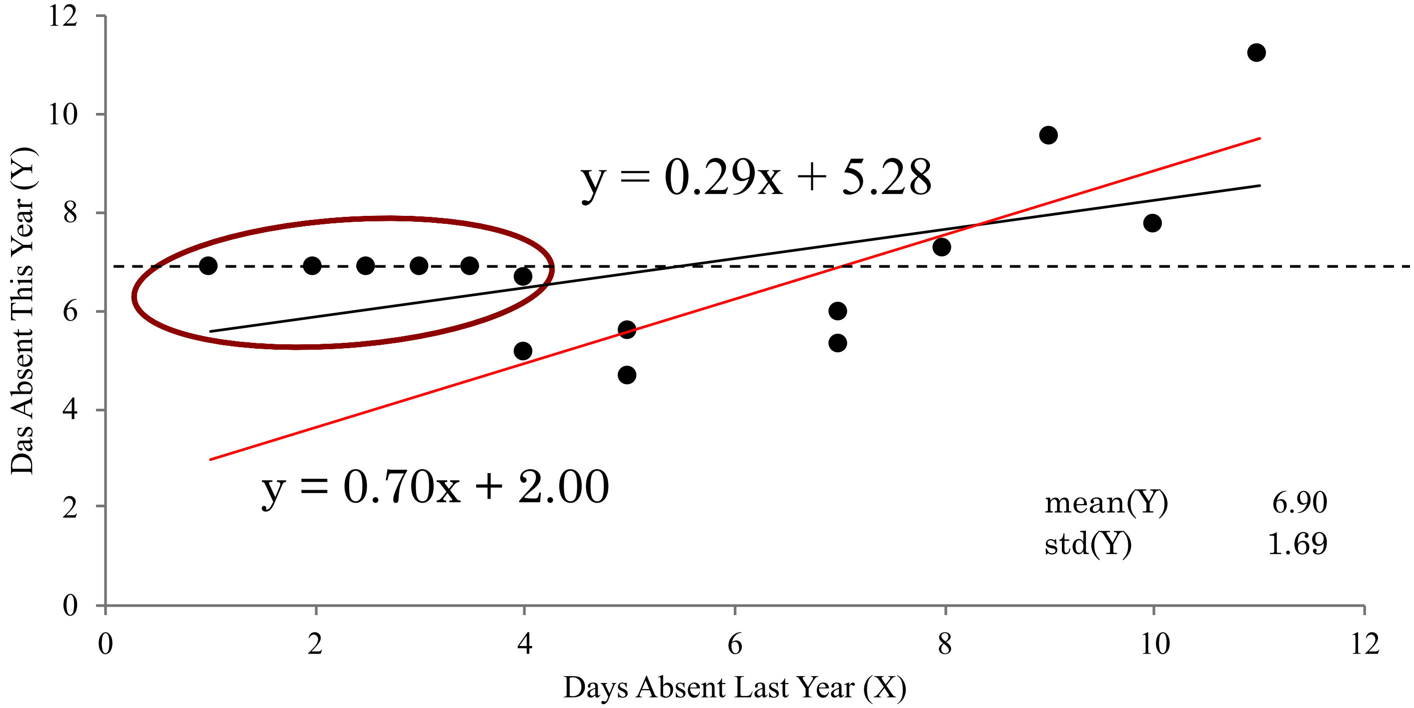
Numerical Example

- Consider this full data set (no missing values)
- Population Intercept=2, Slope= 0.70



Example

- If the lowest 5 values are missing and then mean imputation is used :



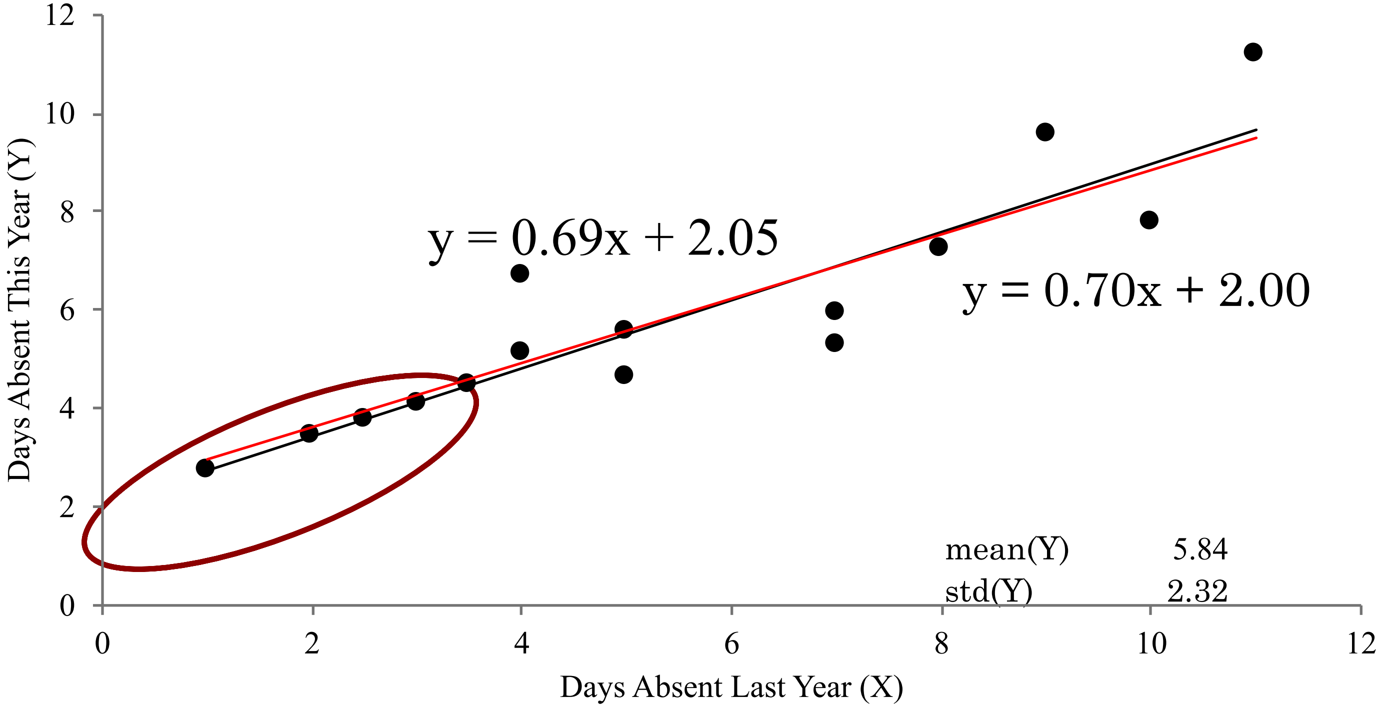
Regression Imputation

- Similar to mean imputation except the missing values are imputed based on predictions from other variables with complete cases in the data

- Better for some estimates (regression coefficients) but variance estimates are still biased

EXAMPLE

- If the lowest 5 values are missing and then regression imputation is used :



MLDS Center Maryland Longitudinal Data System
Better Data • Informed Choices • Improved Results

Summary

- Although traditional methods are easy to implement and, on the surface, appear to address the missing values, they often introduce large biases
  - Particularly with variances
- Two alternative methods are much more preferable for most types of missing data seen in practice
  - Maximum Likelihood
  - Multiple Imputation

Maximum Likelihood

General Overview

- Unlike classical deletion methods or single imputation, maximum likelihood (ML) does not delete observations with missing values or impute values into the raw data

- Instead, ML uses all the observed data for each case.

- Resulting estimates will be unbiased so long as missing values are MAR or MCAR

Background Detail on ML

- ML is an extremely popular method in statistics for estimating model parameters
  - Ex – Estimating the predicted increase in test scores for a treatment.

- Very generally, ML can be thought of as a mathematically sophisticated version of guess and check.

- Each statistical model has a likelihood function that has a mathematical formula
  - Every time we “guess” values, the likelihood function tells us how likely it is that the data came from the proposed estimates

Example

- Suppose the true value of β is 10 (but we don't know this and are trying to estimate it)

$$Test\ Score = \beta \times Treatment$$

- We start with an initial guess of β =1.

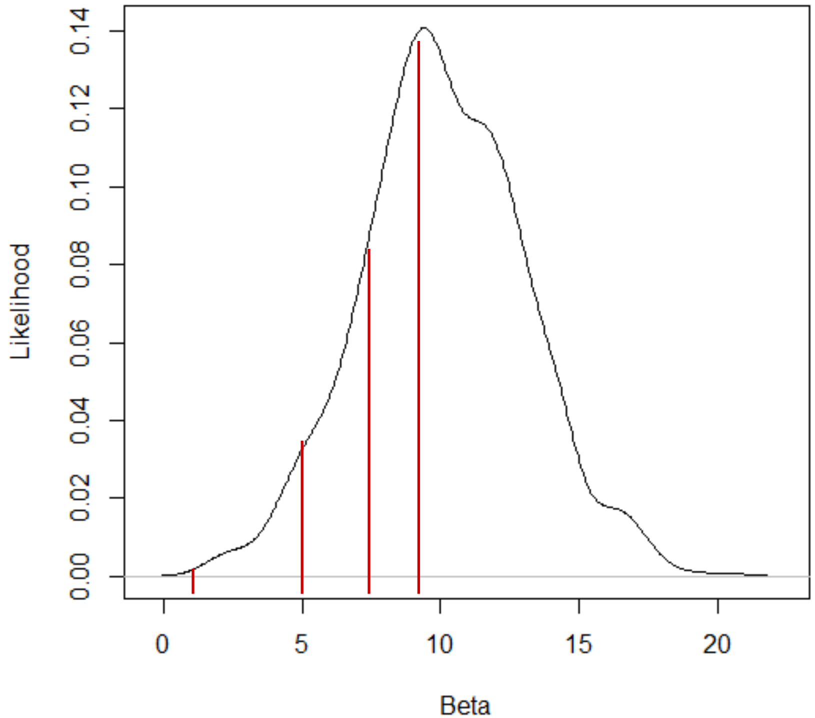
- For some data we have, the likelihood for β =1 is 0.004.
  - The value is not important but higher is better

- We then perform some calculus and come up with a second guess that β =5.8 and the likelihood is now 0.041

Example

- We keep guessing and checking and when we get close to β=10, we should hit the maximum likelihood. We stop here and use this value as our best estimate

  - Graphically, the likelihood looks like a mountain and we are trying to find the estimate for β that gets us to the peak



ML With Missing Data

- Imagine a test is used for placement into an honors program
- Students scoring 90 or less are not admitted to the program and therefore do not have an Honors GPA
- Suppose Student 7 and Student 2 transferred in and do not have a total GPA either

ID	Test Score	Honors GPA	Total GPA
1	78		3.48
2	84		
3	84		3.61
4	85		3.69
5	87		3.73
6	91	3.63	3.81
7	92	3.72	
8	94	3.82	3.98
9	94	3.68	3.90
10	96	3.68	3.95

MLDS Center
Maryland Longitudinal Data System
Better Data • Informed Choices • Improved Results

Using ML to deal with missing data

- Suppose we have a simple research question and just want to know how Test Scores (T), Honors GPA (H), and Total GPA (G) relate to one another.
  - i.e., just want the covariance between scores

$$\text{Covariance} = \begin{bmatrix} Var_T & & \\ Cov_{HT} & Var_H & \\ Cov_{GT} & Cov_{GH} & Var_G \end{bmatrix}$$

MLDS CENTER
Maryland Longitudinal Data System
Better Data • Informed Choices • Improved Results

USING ML TO DEAL WITH MISSING DATA

- The **variance** of each variable measures how spread out the variable is
  - e.g., are GPAs huddle around 3.00 or are some around 2.00 and other around 4.00

$$\text{Covariance} = \begin{bmatrix} Var_T & & \\ Cov_{HT} & Var_H & \\ Cov_{GT} & Cov_{GH} & Var_G \end{bmatrix}$$

USING ML TO DEAL WITH MISSING DATA

- The **covariances** are very similar to a correlation and inform what how variable changes when the other changes
  - e.g., when Total GPA increases, Honors GPA increases too

$$\text{Covariance} = \begin{bmatrix} Var_T & & \\ Cov_{HT} & Var_H & \\ Cov_{GT} & Cov_{GH} & Var_G \end{bmatrix}$$

Using ML to deal with missing data

- Standard estimation methods require that the data are complete with no missing values
  - When the data set is **incomplete**, the estimation process breaks down
- Deletion methods and single imputation try to form a complete data set so standard estimation methods can operate as usual
  - Holes in the data are either removed or filled in

USING ML TO DEAL WITH MISSING DATA

- ML doesn't care if the matrix is complete, it will use whatever information is available
  - ML offers an alternative estimation methods rather than altering the data to fit existing estimation methods
- It does this by changing the size of the matrix for each individual person in the data to extract all possible information.

MLDS CENTER
Maryland Longitudinal Data System
Better Data • Informed Choices • Improved Results

- Student 3 has a Test Score and Total GPA but no Honors GPA
- Rather than forcing the matrix to be complete, ML just works with whatever it is given

$$\begin{bmatrix} Var_T & & \\ Cov_{HT} & Var_H & \\ Cov_{TG} & Cov_{GH} & Var_G \end{bmatrix}$$

$$= \begin{bmatrix} Var_T & \\ Cov_{TG} & Var_G \end{bmatrix}$$

ID	Test Score	Honors GPA	Total GPA
1	78		3.48
2	84		
3	84		3.61
4	85		3.69
5	87		3.73
6	91	3.63	3.81
7	92	3.72	
8	94	3.82	3.98
9	94	3.68	3.90
10	96	3.68	3.95

- Student 2 has a Test Score but no Total GPA or Honors GPA
- With ML, this isn't an issue, it just changes the matrix for Student 2 to get whatever information it can from Student 2

$$\begin{bmatrix} Var_T & & \\ Cov_{HT} & Var_H & \\ Cov_{TG} & Cov_{GH} & Var_G \end{bmatrix}$$

ID	Test Score	Honors GPA	Total GPA
1	78		3.48
2	84		
3	84		3.61
4	85		3.69
5	87		3.73
6	91	3.63	3.81
7	92	3.72	
8	94	3.82	3.98
9	94	3.68	3.90
10	96	3.68	3.95

- Student 9 has all 3 scores
- In this case then the matrix is complete since there are no missing values

$$\begin{bmatrix} Var_T & & \\ Cov_{HT} & Var_H & \\ Cov_{TG} & Cov_{TG} & Var_G \end{bmatrix}$$

ID	Test Score	Honors GPA	Total GPA
1	78		3.48
2	84		
3	84		3.61
4	85		3.69
5	87		3.73
6	91	3.63	3.81
7	92	3.72	
8	94	3.82	3.98
9	94	3.68	3.90
10	96	3.68	3.95

How well does ML actually do?

- <u>With complete data before any values were missing</u>

$$\text{Covariance} = \begin{bmatrix} 30.0 & & \\ .479 & .011 & \\ .864 & .015 & .025 \end{bmatrix}$$

MLDS Center
Maryland Longitudinal Data System
Better Data • Informed Choices • Improved Results

How well does ML actually do?

- With Listwise Deletion (N=4)

$$\text{Covariance} = \begin{bmatrix} 4.25 & & \\ .058 & .007 & \\ .127 & .005 & .006 \end{bmatrix}$$

- With complete data before any values were missing

$$\text{Covariance} = \begin{bmatrix} 30.0 & & \\ .479 & .011 & \\ .864 & .015 & .025 \end{bmatrix}$$

MLDS Center Maryland Longitudinal Data System

Better Data • Informed Choices • Improved Results

How well does ML actually do?

- With Listwise Deletion (N=4)

$$\text{Covariance} = \begin{bmatrix} 4.25 & & \\ .058 & .007 & \\ .127 & .005 & .006 \end{bmatrix}$$

- With complete data before any values were missing

$$\text{Covariance} = \begin{bmatrix} 30.0 & & \\ .479 & .011 & \\ .864 & .015 & .025 \end{bmatrix}$$

- With ML

$$\text{Covariance} = \begin{bmatrix} 29.9 & & \\ .300 & .006 & \\ .834 & .010 & .024 \end{bmatrix}$$

ML improves as sample sizes grow, but even with 10 people it does a much, much better job than listwise deletion.

Why does ML work?

- ML is valid when data are MAR

- The mathematics of ML estimation essentially borrows information from the variables that account for missingness to "makeup" for the missing values
  - Means the variables related to missingness need to be included in the data

- With larger samples, this borrowing makes up for the missing values

Summary of ML

- Because ML uses whatever information is available, it is sometimes referred to as "Full Information Maximum Likelihood"

- It does not impute or delete values, it simply uses whatever data were observed to obtain estimates
  - It uses all the available information each person brings, however much or little

Drawback of Maximum Likelihood

- ML is a useful and legitimate way to handle MAR missing values

- ML does not address the missing values directly but instead works around them and makes the most of the values that are available

- Sometimes it is desirable to have a complete data set without missing values

Multiple Imputation

Overview of Multiple Imputation

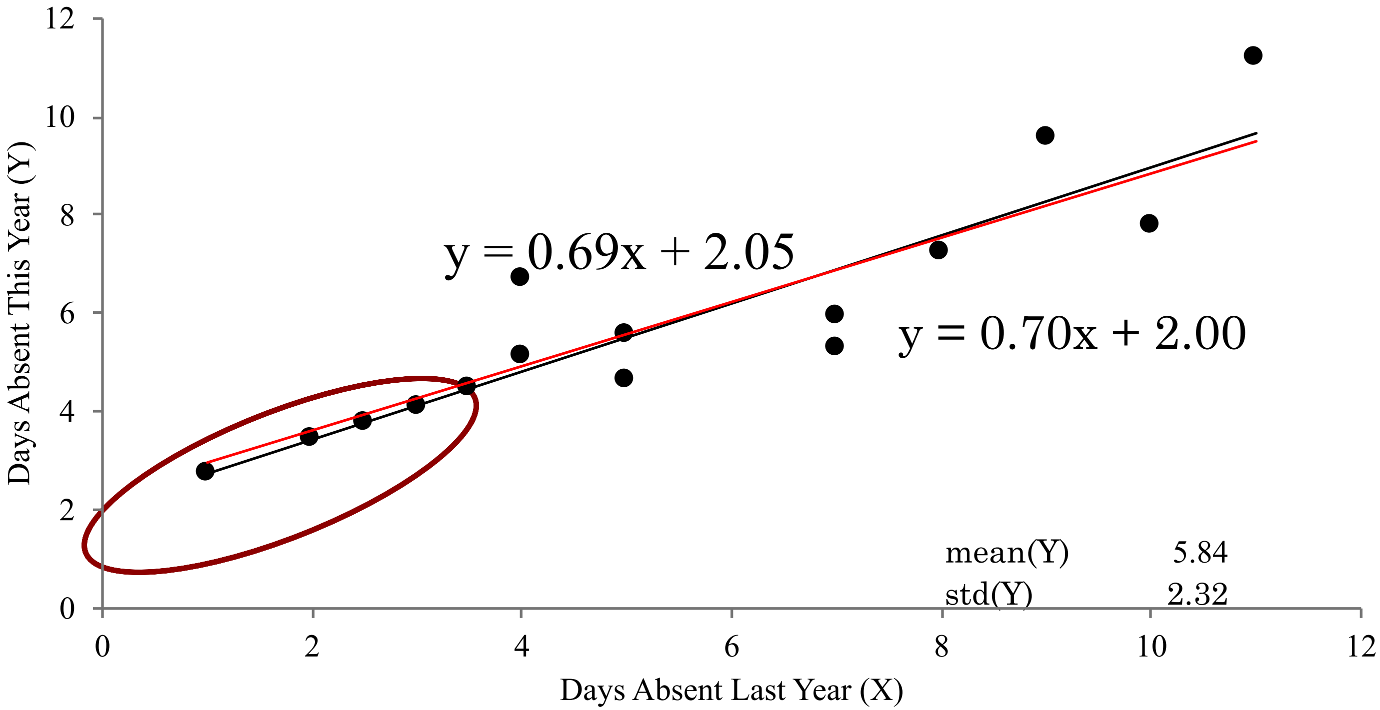
- As discussed earlier, single imputation methods use the data that do have values to fill in the missing values

- Although this seems reasonable, it can create problems because it treats the predicted values as observed
  - Imputed values have prediction error

- As a result, the precision of the estimates is overstated

Recall Regression Imputation

- Discussed how the imputation was a little artificial because the points all fell on the same line

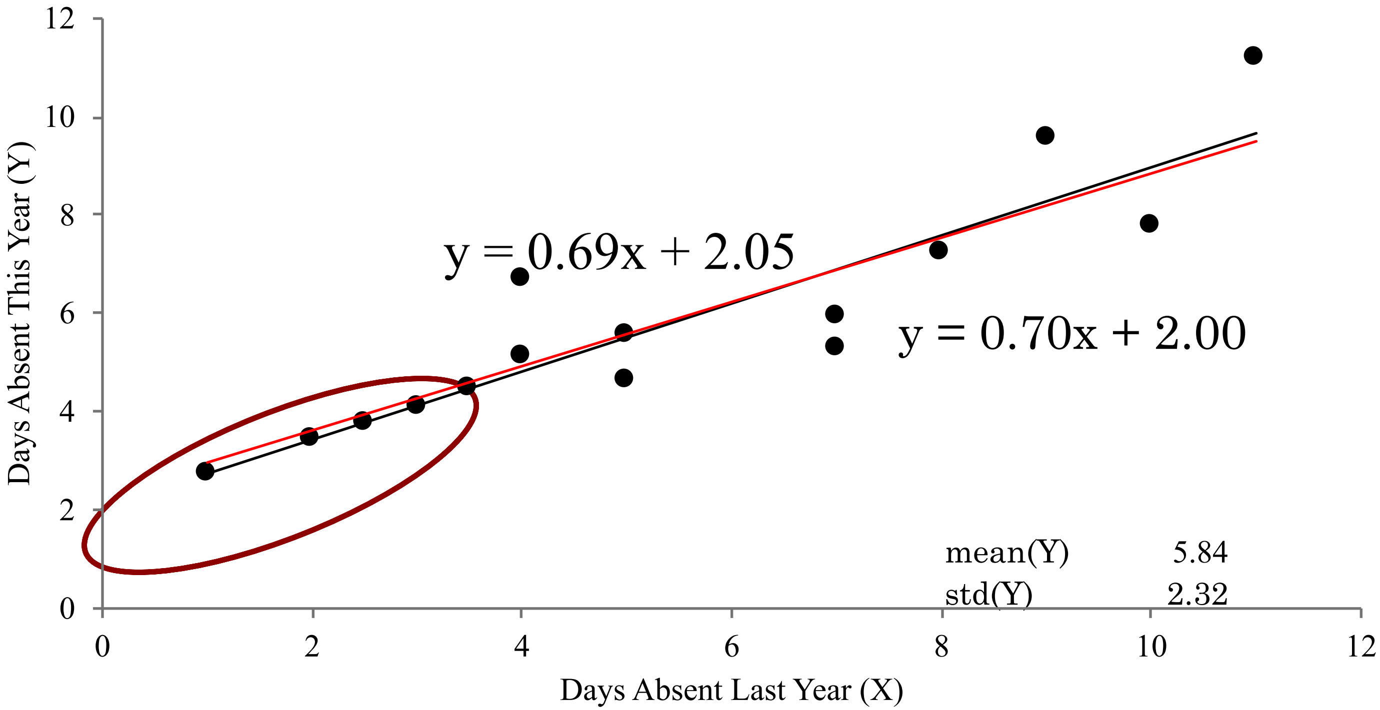


Overview of Multiple Imputation

- Multiple Imputation (MI) addresses this exact problem
- Instead of using the data to fill in the missing values once, MI estimates several plausible values
- Assuming the missing values are MAR, this allows the prediction error for the missing values to be quantified and accounted for in the model

Recall Regression Imputation

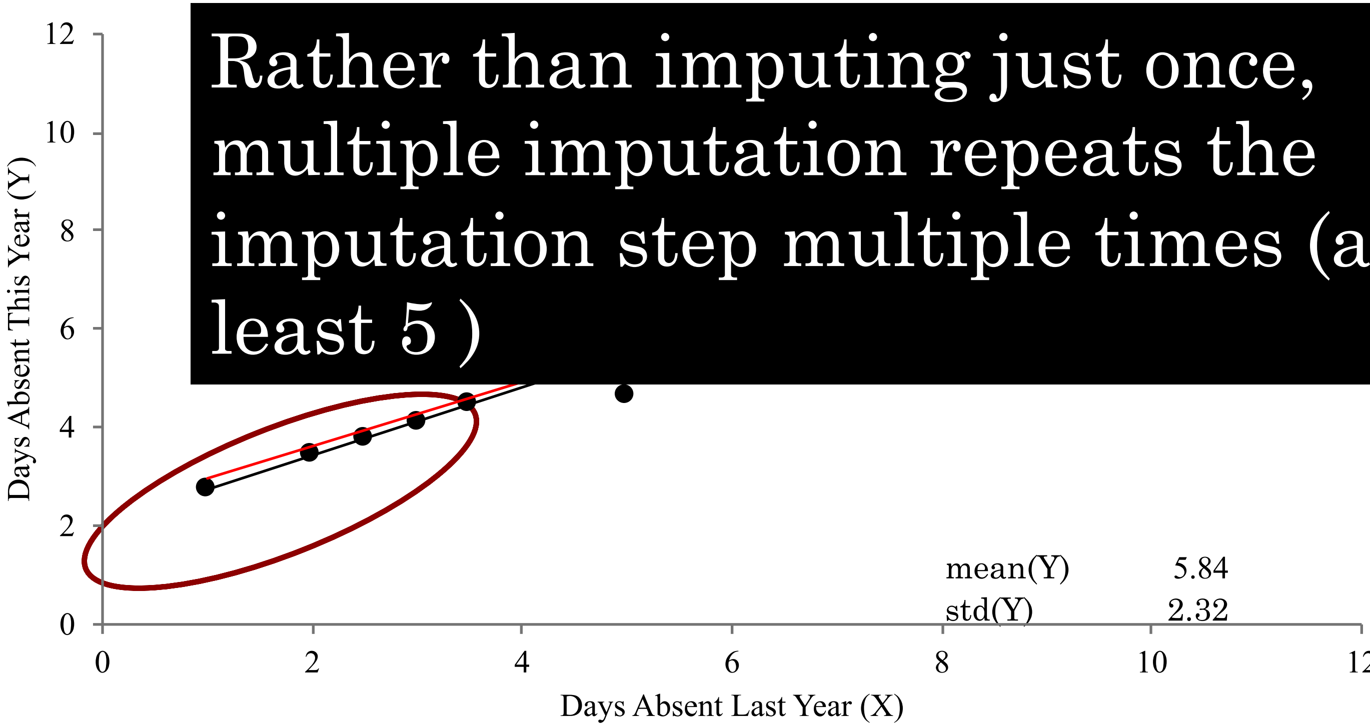
- Discussed how the imputation was a little artificial because the points all fell on the same line



Recall Regression Imputation

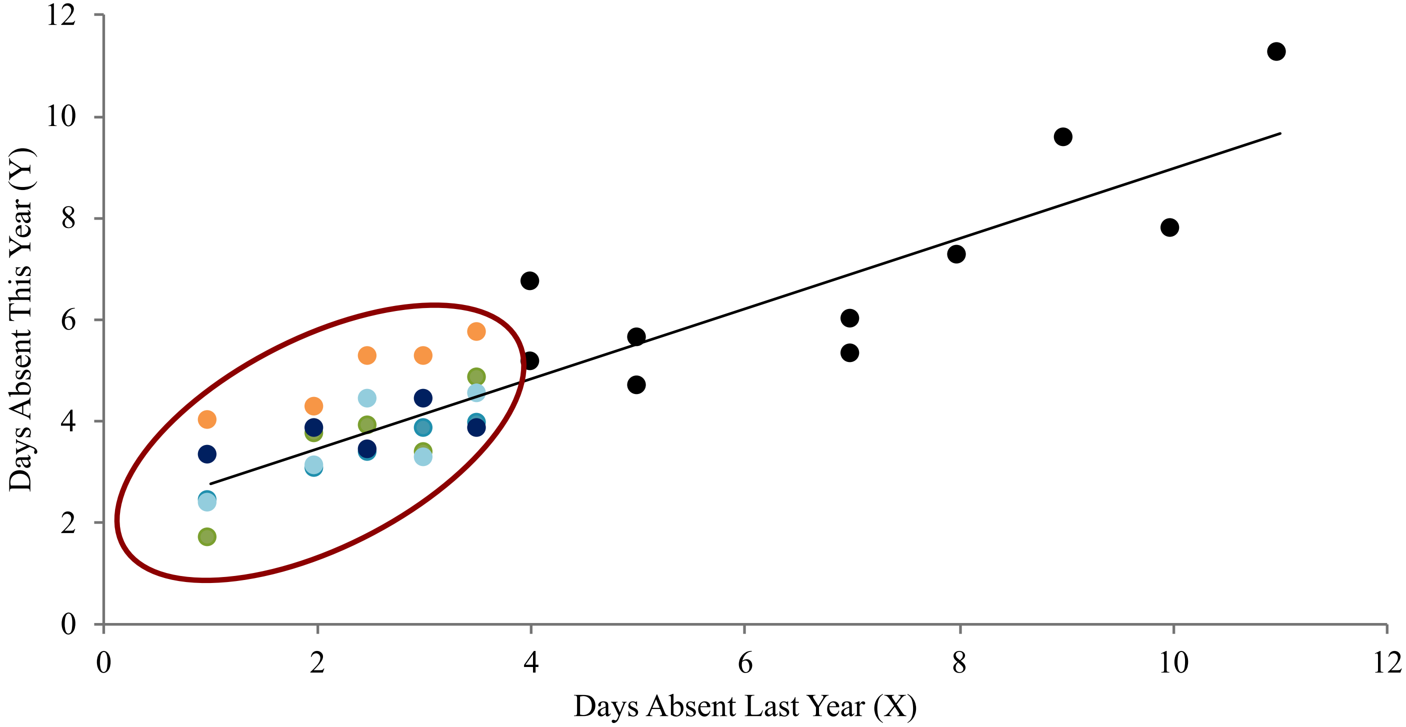
- Discussed how the imputation was a little artificial because the points all fell on the same line

Rather than imputing just once, multiple imputation repeats the imputation step multiple times (at least 5)



Multiple Imputation example

- Multiple imputation makes many copies of the dataset, each one with different imputed values



Multiple Imputation Example

- Each of these copies of the data is then analyzed separately

Imputation #	Intercept	Slope
1	1.58	0.79
2	2.59	0.89
3	1.65	0.69
4	2.33	0.48
5	2.02	0.76

- These leaves us with 5 different sets of estimates though
- We only want 1 set of estimates
  - As if we had complete data

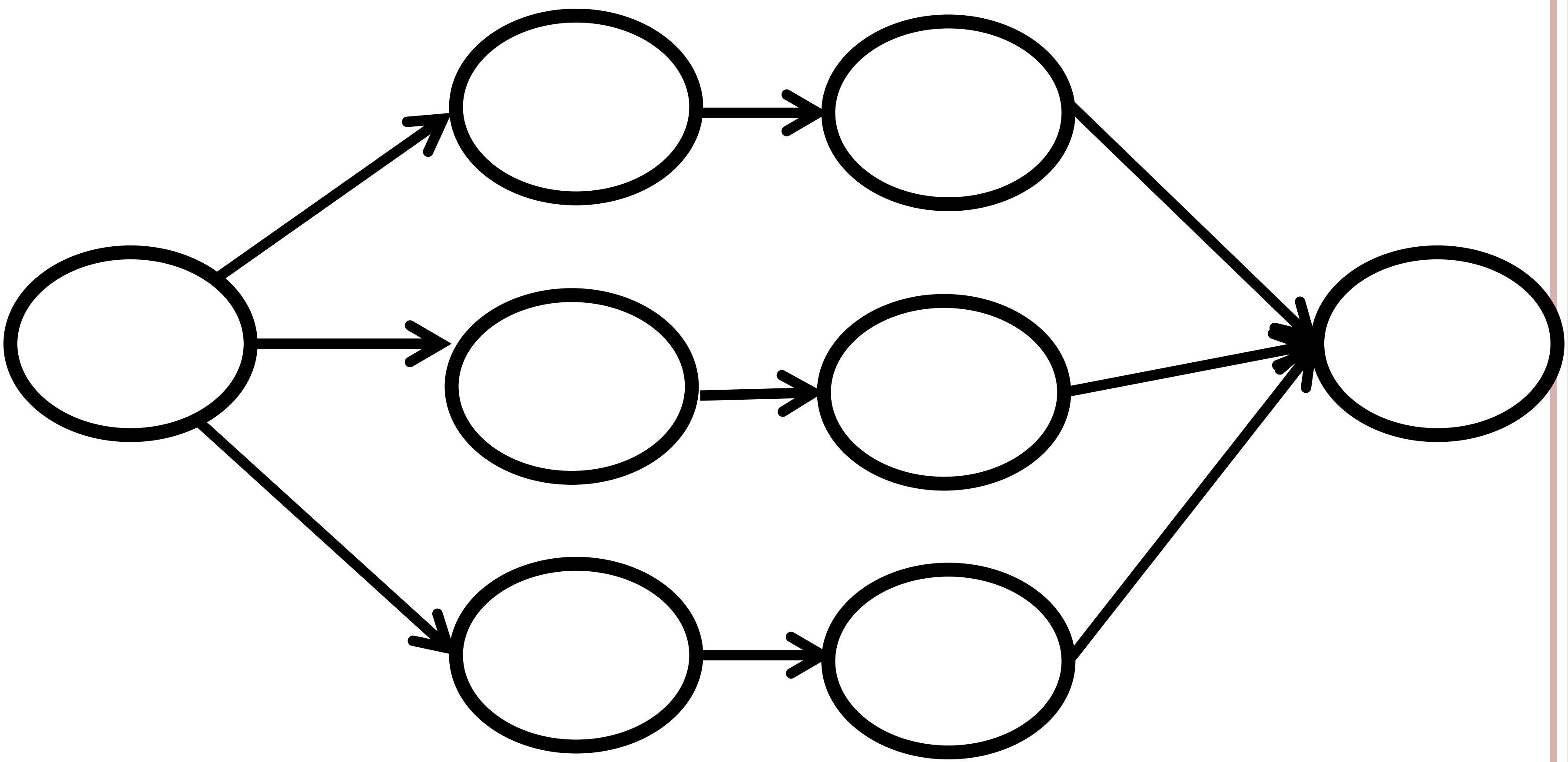
Multiple Imputation Example

- The multiple estimates are then mathematically combined to produced a single set of estimates

Imputation #	Intercept	Slope
1	1.58	0.79
2	2.59	0.89
3	1.65	0.69
4	2.33	0.48
5	2.02	0.76
Average	**2.03**	**0.72**

- Regression coefficients can be combined just by taking the mean
  - Standard errors are much more complicated to combine

Graphical Representation



Data with
Missing Values

Impute
Missing Values

Analyze
Imputed Data

Pool Results

Numerical Example

- Same data set as before for ML example
- Honors GPA is MAR
- Total GPA is MCAR

ID	Test Score	Honors GPA	Total GPA
1	78		3.48
2	84		
3	84		3.61
4	85		3.69
5	87		3.73
6	91	3.63	3.81
7	92	3.72	
8	94	3.82	3.98
9	94	3.68	3.90
10	96	3.68	3.95

First 3 Imputations

ID	Test Score	Honor GPA	Total GPA
1	78	3.51	3.48
2	84	3.55	3.63
3	84	3.52	3.61
4	85	3.61	3.69
5	87	3.63	3.73
6	91	3.63	3.81
7	92	3.72	3.89
8	94	3.82	3.98
9	94	3.68	3.90
10	96	3.68	3.95

ID	Test Score	Honor GPA	Total GPA
1	78	3.56	3.48
2	84	3.61	3.64
3	84	3.56	3.61
4	85	3.65	3.69
5	87	3.65	3.73
6	91	3.63	3.81
7	92	3.72	3.88
8	94	3.82	3.98
9	94	3.68	3.90
10	96	3.68	3.95

ID	Test Score	Honor GPA	Total GPA
1	78	3.57	3.48
2	84	3.65	3.65
3	84	3.57	3.61
4	85	3.66	3.69
5	87	3.64	3.73
6	91	3.63	3.81
7	92	3.72	3.88
8	94	3.82	3.98
9	94	3.68	3.90
10	96	3.68	3.95

Multiple copies of the dataset are created and, in each copy, the missing values have a different predicted value

ID	Score	GPA	GPA
1	78	3.51	3.48
2	84	3.55	3.63
3	84	3.52	3.61
4	85	3.61	3.69
5	87	3.63	3.73
6	91	3.63	3.81
7	92	3.72	3.89
8	94	3.82	3.98
9	94	3.68	3.90
10	96	3.68	3.95

ID	Score	GPA	GPA
1	78	3.56	3.48
2	84	3.61	3.64
3	84	3.56	3.61
4	85	3.65	3.69
5	87	3.65	3.73
6	91	3.63	3.81
7	92	3.72	3.88
8	94	3.82	3.98
9	94	3.68	3.90
10	96	3.68	3.95

			Total
ID	Score	GPA	GPA
1	78	3.57	3.48
2	84	3.65	3.65
3	84	3.57	3.61
4	85	3.66	3.69
5	87	3.64	3.73
6	91	3.63	3.81
7	92	3.72	3.88
8	94	3.82	3.98
9	94	3.68	3.90
10	96	3.68	3.95

MLDS Center
Maryland Longitudinal Data System
Better Data • Informed Choices • Improved Results

Comparison of MI, ML, and Listwise

- With Listwise Deletion

$$\text{Covariance} = \begin{bmatrix} 4.25 & & \\ .058 & .007 & \\ .127 & .005 & .006 \end{bmatrix}$$

- With complete data before any values were missing

$$\text{Covariance} = \begin{bmatrix} 30.0 & & \\ .479 & .011 & \\ .864 & .015 & .025 \end{bmatrix}$$

- With ML

$$\text{Covariance} = \begin{bmatrix} 29.9 & & \\ .300 & .006 & \\ .834 & .010 & .024 \end{bmatrix}$$

Comparison of MI, ML, and Listwise

- With Listwise Deletion

$$\text{Covariance} = \begin{bmatrix} 4.25 & & \\ .058 & .007 & \\ .127 & .005 & .006 \end{bmatrix}$$

- With ML

$$\text{Covariance} = \begin{bmatrix} 29.9 & & \\ .300 & .006 & \\ .834 & .010 & .024 \end{bmatrix}$$

- With complete data before any values were missing

$$\text{Covariance} = \begin{bmatrix} 30.0 & & \\ .479 & .011 & \\ .864 & .015 & .025 \end{bmatrix}$$

- With MI (20 imputations)

$$\text{Covariance} = \begin{bmatrix} 33.3 & & \\ .326 & .007 & \\ .930 & .011 & .026 \end{bmatrix}$$

Summary and Recommendations

Method	Pros	Cons
Listwise Deletion	• Extremely Convenient • Intuitive to explain • Very easy to implement in software	• Primary value is with MCAR data, which isn't seen often in practice • Diminishes sample size very quickly
Maximum Likelihood	• Unbiased with MAR • Treats missing data in one fell swoop • Provides same result every time it is used	• Handles missing data indirectly, no values are deleted or imputed • Difficult to implement without appropriate software
Multiple Imputation	• Unbiased with MAR • Provides a complete dataset • Values can be imputed first and complete data can then be imported to any software program	• Gives different answer each time • Requires an accurate imputation model • Pooling estimates can be challenging

Recommendations for the Center

- Deleting missing observations should generally be avoided if possible since it often will produce biased estimates and greatly reduce sample sizes
  - Could have adverse effects when making policy decisions

- Although variables may not be a direct research interest, they can be important to keep in the data since they can related to missingness
  - Many methods for MAR that are fairly straightforward, MNAR data much more difficult to handle

RECOMMENDATIONS FOR THE CENTER

- Since many different people will be analyzing the same data, it may be helpful to set a policy for how to handle missing values
  - Different methods will give different answers
  - Every instantiation of MI gives different results, so if MI is to be used, it may be useful to include the different plausible values in the data from the onset

- The difficulty with handling missing data also shows that every effort should be made to obtain as much data as possible

References for More Information

- Allison, P. D. (2000). *Missing data*. Thousand Oaks, CA: Sage.
- Enders, C. K. (2010). *Applied missing data analysis*. New York: Guilford Press.
- Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology*, *60*, 549-576.
- Horton, N. J., & Lipsitz, S. R. (2001). Multiple imputation in practice: Comparison of software packages for regression models with missing variables. *American Statistician, 55,* 244–254.
- Little, R. J., & Rubin, D. B. (2002). *Statistical analysis with missing data*. New York: Wiley.
- Rubin, D. B. (1976). Inference and missing data. *Biometrika, 63,* 581–592.
- Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York: Wiley.
- Schafer, J. L., & Graham, J. W. (2002). Missing data: our view of the state of the art. *Psychological Methods*, *7*, 147-177.
- Van Buuren, S. (2012). *Flexible imputation of missing data*. Boca Raton, FL: CRC press.