“An introduction to modern missing data analyses”

Sydney Garcia

Traditional Approaches

1. Exclude cases with missing data (AKA pairwise or listwise deletion)

1. Replace missing values with the mean (AKA single imputation)

Traditional Approaches

1. Exclude cases with missing data (AKA pairwise or listwise deletion)

1. Replace missing values with the mean (AKA single imputation)

Often our data does not meet the assumptions needed!!!

Steps to dealing with missingness

#1 Determine why your data is missing

#2 Select appropriate technique to deal with missingness

Step #1

Determine why your data is missing!

Rubin’s Types of Missing Data

- Missing completely at random (MCAR)
- Missing at random (MAR)
- Missing not at random (MNAR)

(Thank you Donald Rubin!) 🎉

Note

These missing data mechanisms apply to specific analyses

Same dataset can have a mix of Missing completely at random (MCAR), Missing at random (MAR) and Missing not at random (MNAR) variables

Missing Completely at Random (MCAR)

Strict assumption (unlikely to be met in practice)

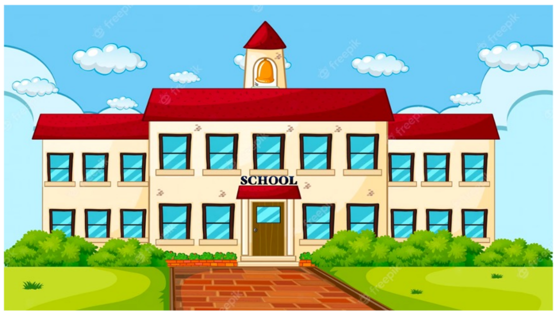
Assumes the probability of missing data on a given variable is unrelated to other variables or to the values of that variable

Missing Completely at Random (MCAR)

Strict assumption (unlikely to be met in practice)

Assumes the probability of missing data on a given variable is unrelated to other variables or to the values of that variable

Basically, probability of being missing is the same for all cases


SCHOOL


SCHOOL

Missing at Random (MAR)

Less strict of an assumption than MCAR– basically, missingness can be related to other measured variables in the analysis model, but not to the underlying values of the incomplete variable

Missing at Random (MAR)

Less strict of an assumption than MCAR– basically, missingness can be related to other measured variables in the analysis model, but not to the underlying values of the incomplete variable

Probability of being missing is the same within groups of observed data

Missing at Random (MAR)

Less strict of an assumption than MCAR– basically, missingness can be related to other measured variables in the analysis model, but not to the underlying values of the incomplete variable

Probability of being missing is the same within groups of observed data

Example: Students can take a class if they pass a test. Missing grades in the class are due to how they did on the test, so need to account for the test when looking at mean grades

Missing Not at Random (MNAR)

MNAR if the probability of missing data is related to the hypothetical values that are missing

Missing Not at Random (MNAR)

MNAR if the probability of missing data is related to the hypothetical values that are missing

Or if we don’t know why the data are missing

Missing Not at Random (MNAR)

MNAR if the probability of missing data is related to the hypothetical values that are missing

Example: On reading test, poor readers may fail to respond to some questions. So the probability of missingness is directly related to reading ability (that we are trying to measure), but we probably didn’t know its the questions themselves that are leading to missingness!

Problem

Can only test for MCAR because MAR and MNAR depend on **unobserved data**

Mostly people assume MAR

Problem

Can only test for MCAR because MAR and MNAR depend on **unobserved data**

Mostly people assume MAR

Note, again: each variable could be missing for a different reason!

Deciding type of missingness

To know if its MCAR or MAR, could do followup survey with missing participants. If not too different from rest of the group, then probably MCAR

Deciding type of missingness

To know if its MCAR or MAR, could do followup survey with missing participants. If not too different from rest of the group, then probably MCAR

If can't do followup, can test if missingness is related to any variables in the analysis using R packages

Step #2

Select appropriate estimation method

MAR

MCAR

MNAR

Pretty open!

Use multiple imputation or maximum likelihood estimation

Stochastic regression imputation also possible

Use multiple imputation or maximum likelihood estimation

Though estimates will still be biased– research still being done

What happens when you use the wrong method?

What happens when you use the wrong method on non-MCAR data?

List-wise deletion (incomplete cases removed)

- Reduce sample size → reduce power of significance tests
- produces biased estimates

Pairwise deletion (incomplete cases removed analysis-by-analysis)

- Helps preserve some power
- produces biased estimates

Example of incorrectly using deletion

Table 1
Math performance data set.

Complete data		Observed data	Mean imputation	Regression imputation[a]	Stochastic regression imputation[a]	
Math aptitude	Course grade	Course grade	Course grade	Course grade	Random error	Course grade
4.0	71.00	–	81.80	65.26	7.16	72.42
4.6	87.00	–	81.80	68.22	0.73	68.95
4.6	74.00	–	81.80	68.22	12.01	80.23
4.7	67.00	–	81.80	68.71	−7.91	60.81
4.9	63.00	–	81.80	69.70	−4.07	65.63
5.3	63.00	–	81.80	71.68	27.41	99.09
5.4	71.00	–	81.80	72.17	25.76	97.93
5.6	71.00	–	81.80	73.16	2.76	75.92
5.6	79.00	–	81.80	73.16	−11.77	61.39
5.8	63.00	–	81.80	74.15	−0.56	73.59
6.1	63.00	63.00	63.00	63.00	–	63.00
6.7	75.00	75.00	75.00	75.00	–	75.00
6.7	79.00	79.00	79.00	79.00	–	79.00
6.8	95.00	95.00	95.00	95.00	–	95.00
7.0	75.00	75.00	75.00	75.00	–	75.00
7.4	75.00	75.00	75.00	75.00	–	75.00
7.5	83.00	83.00	83.00	83.00	–	83.00
7.7	91.00	91.00	91.00	91.00	–	91.00
8.0	99.00	99.00	99.00	99.00	–	99.00
9.6	83.00	83.00	83.00	83.00	–	83.00
Mean	76.35	81.80	81.80	76.12		78.70
Std. Dev.	10.73	10.84	7.46	9.67		12.36

[a] Imputation regression equation: $\hat{Y}=45.506+4.938(\text{Aptitude})$.

If only looking at complete cases….

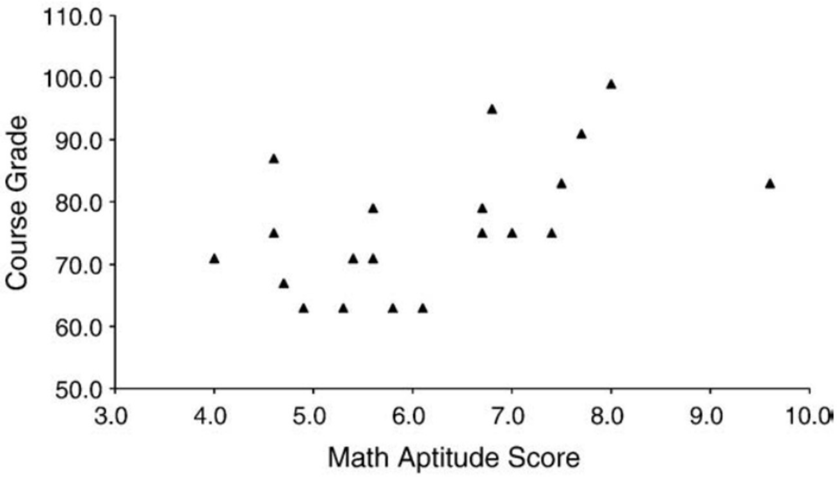


Fig. 1. Complete-data scatterplot of the math performance data in Table 1.

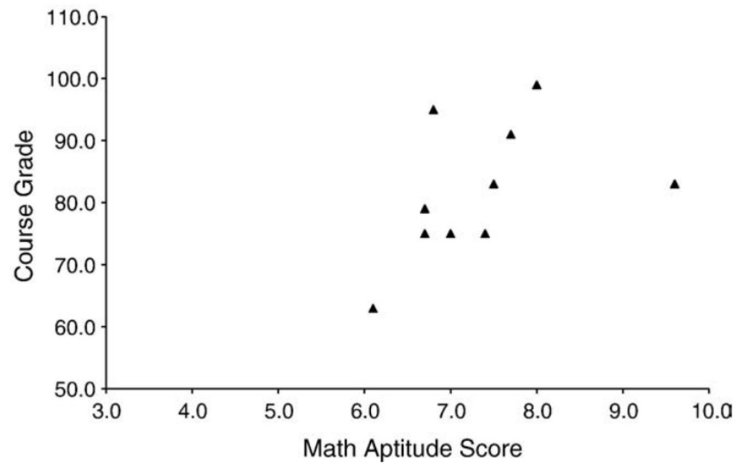


Fig. 2. Listwise deletion scatterplot of the math performance data in Table 1.

Problem with deletion

We chopped off the lower part of the distribution for these variables, so the means are too high and estimates of variability are too low!

Another wrong method: single imputation

For example, replace missing values with the mean, or use predicted values from the regression equation

This will reduce variability in the data and give us an incorrect correlation!

Another wrong method: single imputation

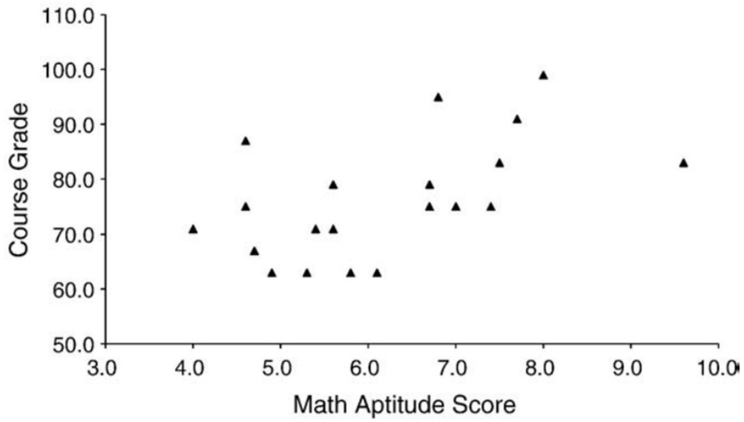


Fig. 1. Complete-data scatterplot of the math performance data in Table 1.

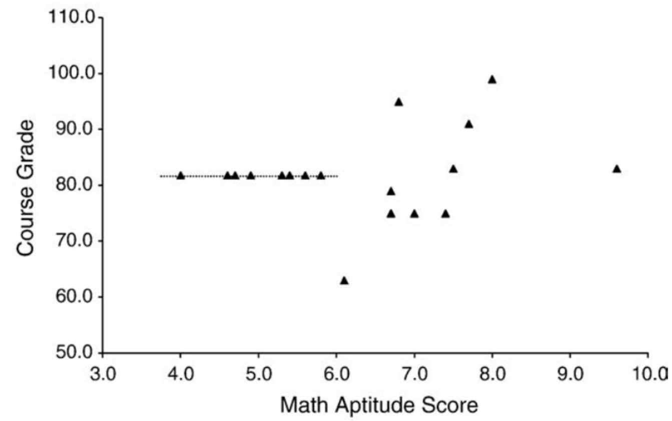


Fig. 3. Mean imputation scatterplot of the math performance data in Table 1.

TLDR

Mean or regression imputation → bias because does not account for variability of the hypothetical values

Stochastic regression imputation → **better** because adds random error to the predicted values from regression imputation

TLDR

Mean or regression imputation → bias because does not account for variability of the hypothetical values

Stochastic regression imputation → **better** because adds random error to the predicted values from regression imputation

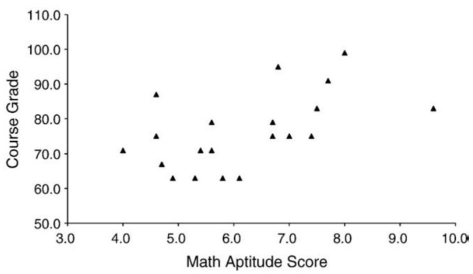


Fig. 1. Complete-data scatterplot of the math performance data in Table 1.

vs.

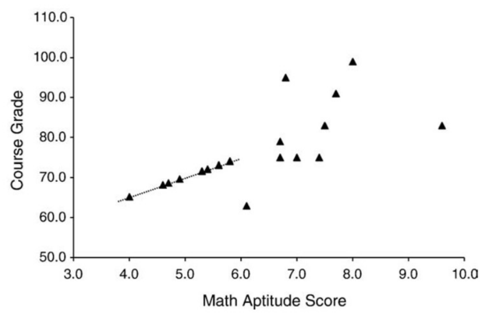


Fig. 4. Regression imputation scatterplot of the math performance data in Table 1.

vs.

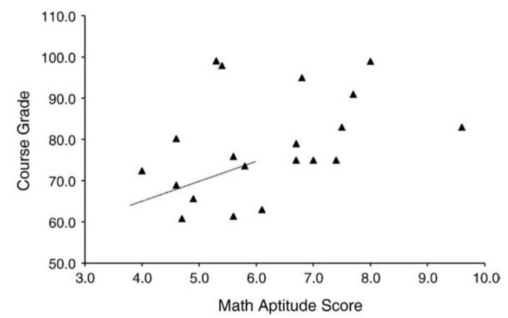


Fig. 5. Stochastic regression imputation scatterplot of the math performance data in Table 1.

One preferable method for MAR data that is normal

Multiple Imputation

One preferable method for MAR data that is normal

Multiple Imputation

1. Impute data
   a. Generate many data sets (20 is recommended) with different estimates of the missing values
2. Analyze data
   a. Get multiple parameter estimates and standard errors
3. Pool results
   a. Combine all results

Imputing Data

Step 1: Imputation step, similar to stochastic regression, you use regression equation to predict the incomplete variables from complete variables, and add a normally distributed residual term to add variability to data

Imputing Data

Step 1: Imputation step, similar to stochastic regression, you use regression equation to predict the incomplete variables from complete variables, and add a normally distributed residual term to add variability to data

Step 2: Use Bayesian estimation to generate new estimates of means and covariances and add a random residual term to each of these estimates (so that these values randomly differ)

Imputing Data

Step 1: Imputation step, similar to stochastic regression, you use regression equation to predict the incomplete variables from complete variables, and add a normally distributed residual term to add variability to data

Step 2: Use Bayesian estimation to generate new estimates of means and covariances and add a random residual term to each of these estimates (so that these values randomly differ)

Then use updated parameter estimates to construct new set of imputations which differ from previous 2 steps. Do these steps many times (with iterations in between so that the data sets are independent)

Table 2
Imputed course grades from multiple imputation procedure.

Observed data		Imputed course grades			
Math aptitude	Course grade	Data Set 1 [a]	Data Set 2 [b]	Data Set 3 [c]	Data Set 4 [d]
4.00	–	51.48	67.91	69.38	72.45
4.60	–	59.53	62.59	74.19	57.38
4.60	–	62.34	59.77	67.43	46.47
4.70	–	68.45	53.56	71.39	56.99
4.90	–	75.47	63.79	72.54	85.96
5.30	–	81.81	57.16	70.99	68.71
5.40	–	61.05	90.47	56.25	74.11
5.60	–	77.72	46.92	69.14	52.91
5.60	–	71.49	70.79	73.89	72.44
5.80	–	68.36	59.98	67.04	77.53
6.10	63.00	63.00	63.00	63.00	63.00
6.70	75.00	75.00	75.00	75.00	75.00
6.70	79.00	79.00	79.00	79.00	79.00
6.80	95.00	95.00	95.00	95.00	95.00
7.00	75.00	75.00	75.00	75.00	75.00
7.40	75.00	75.00	75.00	75.00	75.00
7.50	83.00	83.00	83.00	83.00	83.00
7.70	91.00	91.00	91.00	91.00	91.00
8.00	99.00	99.00	99.00	99.00	99.00
9.60	83.00	83.00	83.00	83.00	83.00
Mean	81.80	74.79	72.55	75.51	74.15
SE	10.84	12.18	14.53	10.49	13.81

[a] Imputation regression equation: $\hat{Y}=6.03(\text{Aptitude})+33.92$.
[b] Imputation regression equation: $\hat{Y}=5.11(\text{Aptitude})+38.49$.
[c] Imputation regression equation: $\hat{Y}=5.62(\text{Aptitude})+41.15$.
[d] Imputation regression equation: $\hat{Y}=6.40(\text{Aptitude})+31.54$.

Analyzing Data

Analyze each data set as you normally would and get multiple estimates of the parameter and standard errors

Pooling data

Get pooled parameter estimates by taking the mean of all the estimates

Get pooled standard error

- Need to account for because it involves the standard errors from the imputed data sets (i.e., within-imputation variance) and the extent to which the estimates vary across data sets (i.e., between-imputation variance)

Pooled SE

$$W = \frac{\sum SE_t^2}{m}, \qquad B = \frac{\sum (\hat{\theta}_t - \overline{\theta})^2}{m-1},$$

$$SE = \sqrt{W + B + B/m.}$$

Pooled SE

$$W = \frac{\sum SE_t^2}{m}, \qquad B = \frac{\sum(\hat{\theta}_t - \overline{\theta})^2}{m-1},$$

$$SE = \sqrt{W + B + B/m.}$$

But there are packages to do this for you!

Eg mice package in r

Another correct method for MAR data that is normal

Maximum likelihood estimation

Does not fill in values. Instead, uses existing values to identify the parameter values that have the highest probability of producing the sample data

Another correct method for MAR data that is normal

Maximum likelihood estimation

Does not fill in values. Instead, uses existing values to identify the parameter values that have the highest probability of producing the sample data

Uses a mathematical function called a log likelihood to quantify the standardized distance between the observed data points and the parameters of interest (e.g., the mean), and **the goal is to identify parameter estimates that minimize these distances (like least squares)**

Log likelihood equation

$$logL = \sum_{i=1}^{N} log\left[\frac{1}{\sqrt{2\pi\sigma^2}}e^{-.5\left(\frac{y_i-\mu}{\sigma}\right)^2}\right].$$

Probability density function for shape of normal curve

This is the relative probability of obtaining a single
score from a normally distributed population with a particular (unknown) mean and
standard deviation and score

Log likelihood equation

$$logL = \sum_{i=1}^{N} log\left[\frac{1}{\sqrt{2\pi\sigma^2}} e^{-.5\left(\frac{y_i-\mu}{\sigma}\right)^2}\right].$$

Probability density function for shape of normal curve

This is the relative probability of obtaining a single score from a normally distributed population with a particular (unknown) mean and standard deviation and score

Substitute parameter values (mean/SD) and observed y values to get the standardized distance between that data point and the mean

Log likelihood equation

$$logL = \sum_{i=1}^{N} log\left[\frac{1}{\sqrt{2\pi\sigma^2}} e^{-.5\left(\frac{y_i-\mu}{\sigma}\right)^2}\right].$$

Probability density function for shape of normal curve

This is the relative probability of obtaining a single
score from a normally distributed population with a particular (unknown) mean and standard deviation

adds the relative probabilities into a summary measure (the sample log likelihood)

Substitute parameter values (mean/SD) and observed y values to get the standardized distance between that data point and the mean

Summary

Need to know what kind of missingness you have (MCAR, MAR or MNAR)

Then use either multiple imputation (to estimate missing values) or maximum likelihood estimation (to estimate that parameter values that may have produced that data)

Questions

Do same methods apply for missing data in experiments?

What about when data is not normal?