Title
Show Me the Missing Data

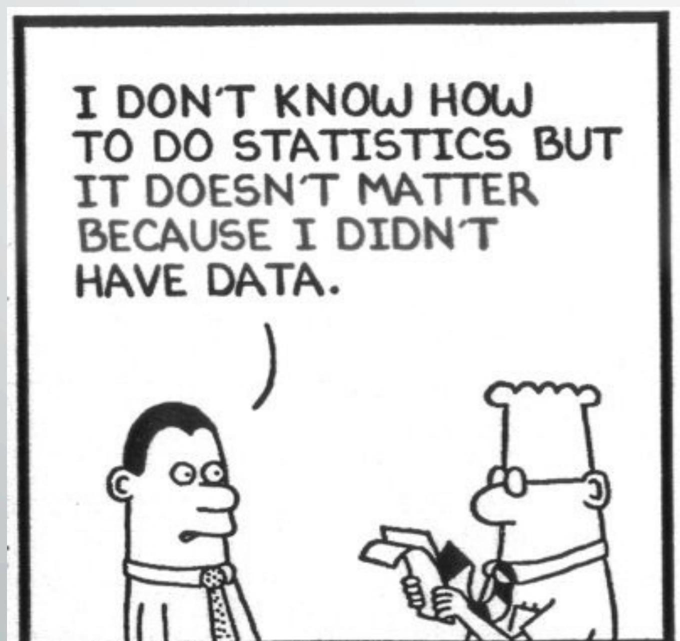
Author
Sanchez, Juana

Publication Date
2017-05-19

Show Me the Missing Data

Juana Sanchez
(with Dennis Li)
UCLA, Department of Statistics

(USCOTS 2017, BOS 2H, 5/19/2017)



Source: https://ducttapefordata.com/category/analysis/

Source: https://www.cartoonstock.com/directory/d/de

...u teach intro stats classes for undergraduates students that never took statistics before?

...ce in a while, not every year, in college

Every year , college

Never, college

...e in a while, not every year, high school

Never, high school

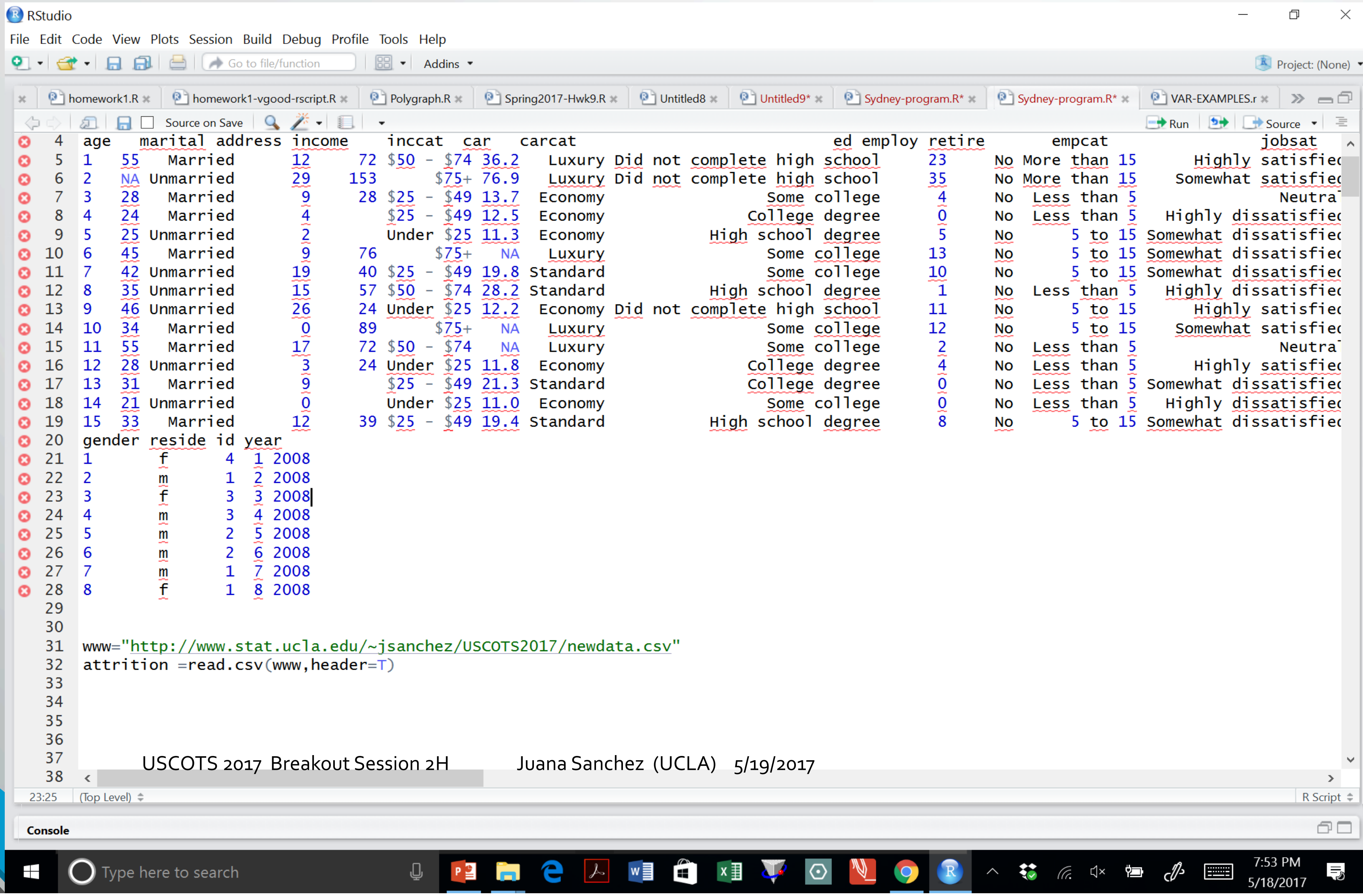
Every year, high school

Start the presentation to activate live content

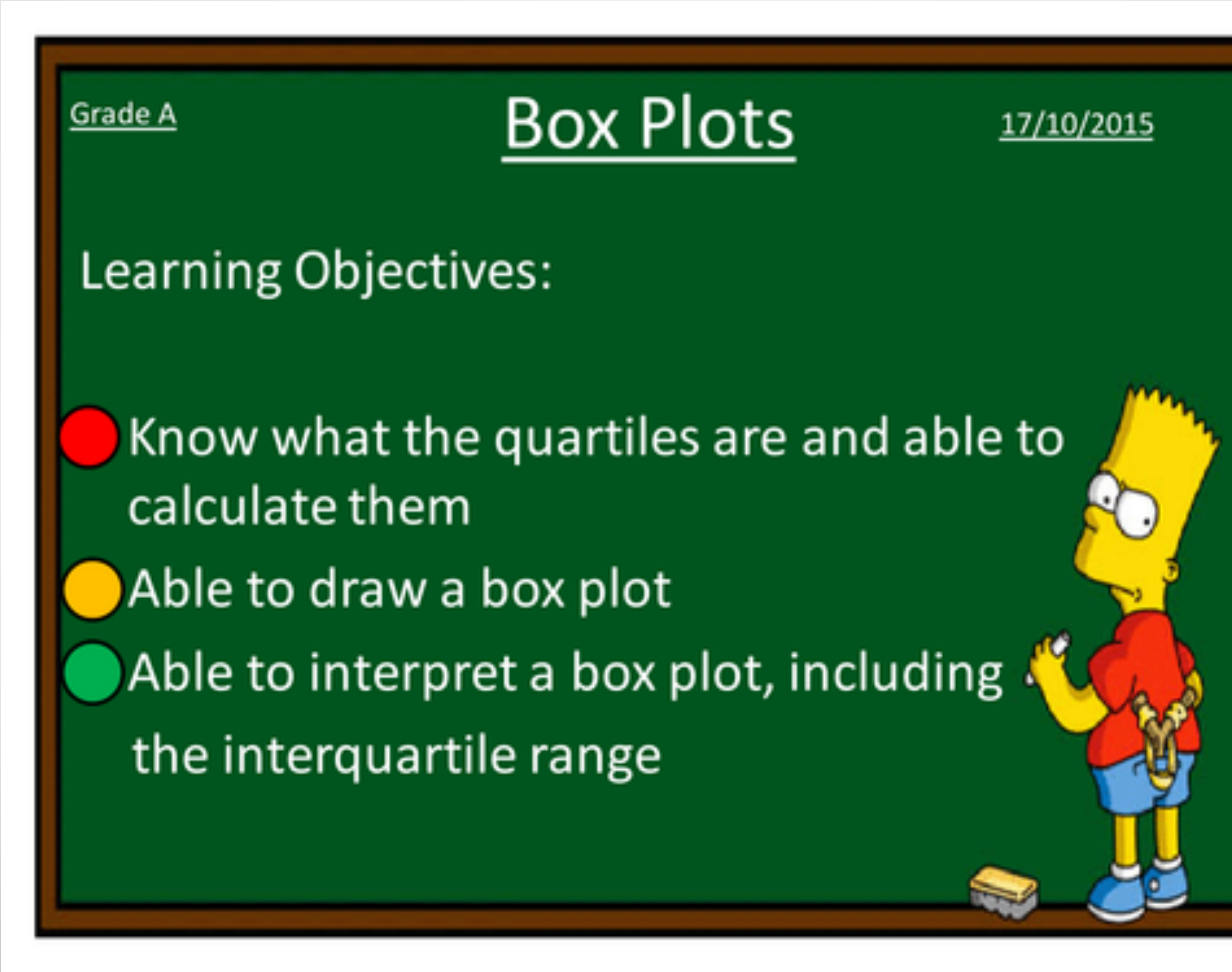
If you see this message in presentation mode, install the add-in or get help at PollEv.com/app

...where

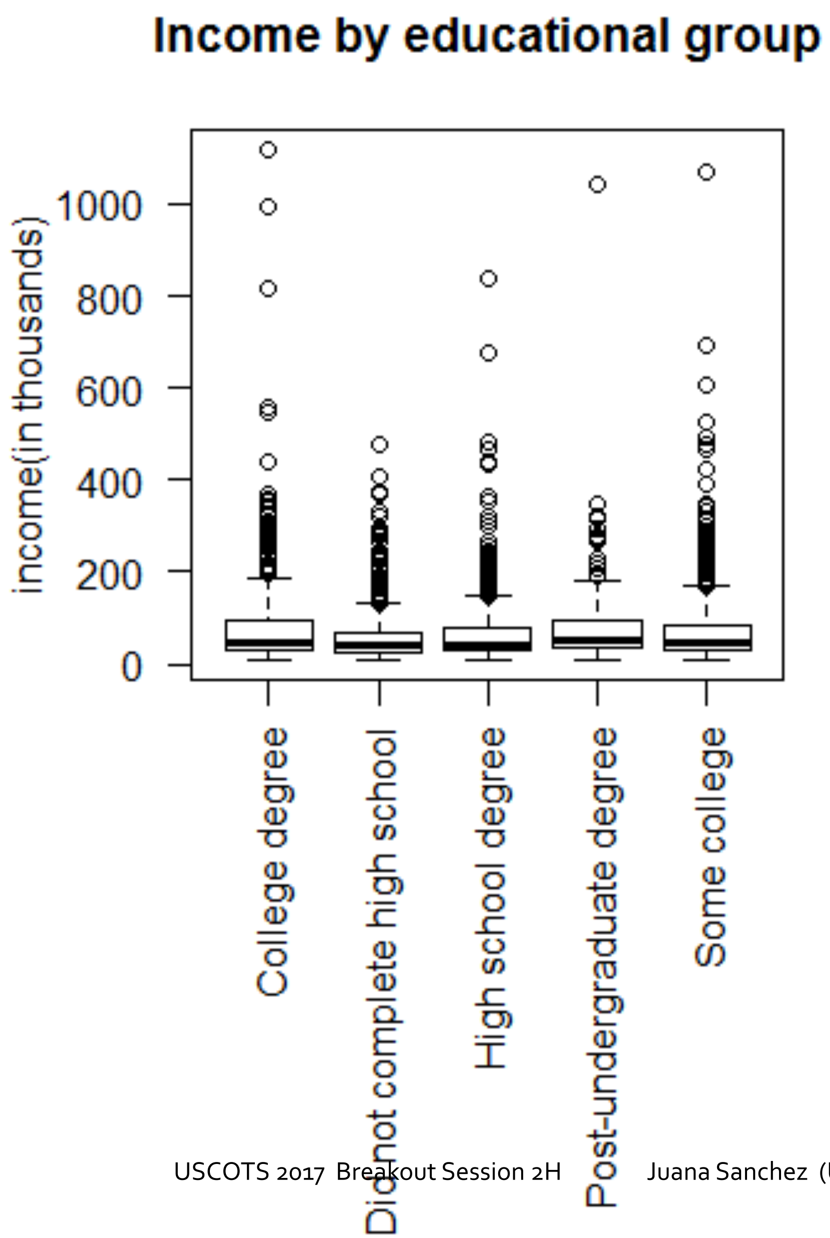
We indulge in data since the first lectures and labs in an intro stats class, e.g., n=3165 and 15 variables

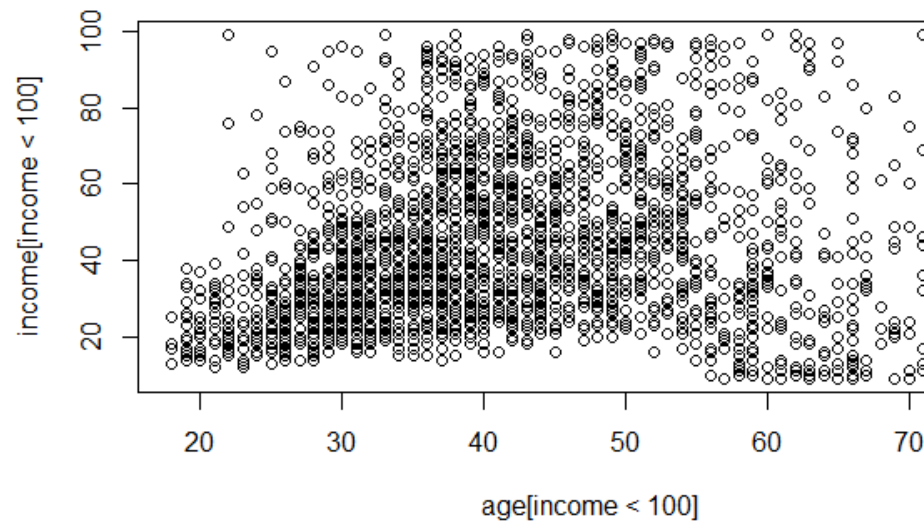


In a decent Stats class we may use that data to achieve, for example, the third of the following learning objectives:



And we get more multivariate later, using the same data.





What do we do with the missing data?

```
age  marital address income   inccat  car  carcat                  ed employ retire      empcat               jobsat
1  55   Married      12    72 $50 - $74 36.2   Luxury Did not complete high school     23    No More than 15     Highly satisfiec
2  NA Unmarried      29   153     $75+ 76.9   Luxury Did not complete high school     35    No More than 15  Somewhat satisfiec
3  28   Married       9    28 $25 - $49 13.7  Economy              Some college      4    No  Less than 5           Neutra
4  24   Married       4       $25 - $49 12.5  Economy            College degree      0    No  Less than 5   Highly dissatisfiec
5  25 Unmarried       2       Under $25 11.3  Economy        High school degree      5    No      5 to 15 Somewhat dissatisfiec
6  45   Married       9    76     $75+   NA   Luxury              Some college     13    No      5 to 15 Somewhat dissatisfiec
7  42 Unmarried      19    40 $25 - $49 19.8 Standard              Some college     10    No      5 to 15 Somewhat dissatisfiec
8  35 Unmarried      15    57 $50 - $74 28.2 Standard        High school degree      1    No  Less than 5   Highly dissatisfiec
9  46 Unmarried      26    24 Under $25 12.2  Economy Did not complete high school     11    No      5 to 15     Highly satisfiec
10 34   Married       0    89     $75+   NA   Luxury              Some college     12    No      5 to 15  Somewhat satisfiec
11 55   Married      17    72 $50 - $74   NA   Luxury              Some college      2    No  Less than 5           Neutra
12 28 Unmarried       3    24 Under $25 11.8  Economy            College degree      4    No  Less than 5     Highly satisfiec
13 31   Married       9       $25 - $49 21.3 Standard            College degree      0    No  Less than 5 Somewhat dissatisfiec
14 21 Unmarried       0       Under $25 11.0  Economy              Some college      0    No  Less than 5   Highly dissatisfiec
15 33   Married      12    39 $25 - $49 19.4 Standard        High school degree      8    No      5 to 15 Somewhat dissatisfiec
gender reside id year
1        f      4  1 2008
2        m      1  2 2008
3        f      3  3 2008
4        m      3  4 2008
5        m      2  5 2008
6        m      2  6 2008
7        m      1  7 2008
8        f      1  8 2008


www="http://www.stat.ucla.edu/~jsanchez/USCOTS2017/newdata.csv"
attrition =read.csv(www,header=T)
```

How do you approach missing data in your intro stats class?

up

Start the presentation to activate live content

If you see this message in presentation mode, install the add-in or get help at PollEv.com/app

where

Why do these things get forgotten so quickly after the intro stats course?

Our students come to our classes with mental model[s] prior knowledge, beliefs and categorizations that ma[y] interfere with the transition from novice to experts i[n] statistics.

Transformed courses have:

- Pre-lecture assignments
- Just-in-time teaching (clickers, in-class polls)
- Socially mediated, collaborative learning, think-pair-share, dialogue.
- In-class tutorials that engage students in coversations between themselves and with the teacher.
- Emphasis on the process by which students learn.

Transform to engage studen[ts]

- Cognitively and emotionally
  - In knowledge construction, via
    - Critical thinking skills ,
    - Long term memory retention,
    - conceptual change., awareness of contradiction
    - Thinking like scientists
    - Talking about what they think.

What about a more investigative ro
What can the data tell us about t
missing da

Hands on acti

Think-pair-share (5 min

the margin plots from VIM package in R telling us about the missin of income, age, and car value (car)?

op

Start the presentation to activate live content

If you see this message in presentation mode, install the add-in or get help at PollEv.com/app

here

What can the data tell us about t
missing da

Hands on acti

Think-pair-share (5 min

Missing Data Techniques: Mechanisms and Methods

Juana Sanchez

Dennis Li

UCLA, Department of Statistics

Missing Data Mechanisms

Mechanisms describe the assumptions about the nature of the missing data and can be categorized as follows:

1. MCAR (Missing Completely at Random)
2. MAR (Missing at Random)
3. NMAR (Not Missing at Random)

MCAR (Missing Completely at Random)

- Probability of missing values has nothing to do with the observed or missin
values
- Example: someone flips a coin to determine if they will fill out a survey, or a
researcher is unable to gather data for a day due to an assay failure

MAR (Missing at Random)

- Probability of missing values depends only on the observed values in the dataset (not the missing variable itself)
- Example: women may be more likely to answer or decline to answer certai questions than men. Older subjects may be more likely to drop out of a study (Gender and age are observed here)

NMAR (Not Missing at Random)

- Probability of missing values depends on the missing values themselves, and can also depend on observed values too
- Example: a study that measures weight has 2 rounds, and people don't sho up to the second round because they've gained weight and believe the stu isn't helping them (missing weight data is due to weight itself)

Some Methods to Handle Missing Data

1. Complete Case (CC) Analysis
2. Inverse Probability Weighting (IPW)
3. Last Observation Carried Forward (LOCF) Imputation
4. Unconditional Mean Imputation
5. Single/Deterministic/Regression Imputation
6. Stochastic Imputation
7. Multiple Imputation

Complete Case Analysis (CC)

- Default method in statistical software packages such as R, Stata, SAS
- Delete whole row which contains missing data on any variable
- **Advantages**: easiest, default, unbiased with MCAR
- **Disadvantages**: loss of valuable data, mostly biased (MCAR is rarest)

Subject	Weight	Age	Sex
1	150	60	F
2	.	43	M
3	190	20	M
4	210	38	M
5	.	19	F

→

Subject	Weight	Age	Sex
1	150	60	F
3	190	20	M
4	210	38	M

Megan M. Marron & Abdus S. Wahed (2016) Teaching Missing Data Method
Undergraduates Using a Group-Based Project Within a Six-Week Summer P
Statistics Education, 24:1, 8-15

Inverse Probability Weighting (IPW)

- Look for similarities between subjects who are missing the outcome of interest vs those who are not
- Find pairings where similarities exist, and calculate the probability of missing the outcome of interest based on pairings
- Assumes MAR data (allows calculation to be based on observed info)
- **Advantages**: results are unbiased under MAR and MCAR
- **Disadvantages**: reduced sample size, skewed if small predicted probability of complete data

Example of IPW

Table 1. Data used to explain IPW.

Subject	Age	Sex	Year in College
1	.	F	Graduated
2	.	F	Junior
3	20	M	Junior
4	24	F	Graduated
5	21	F	Senior
6	19	F	Junior

$$\text{Estimated Mean Age} = \frac{1}{6}(\text{Subject3}'\text{sage} + 2 * \text{Subject4}'\text{sage} + \text{Subject5}'\text{sage} + 2 * \text{Subject6}'\text{sage})$$

$$= \frac{1}{6}\left(\frac{Y_{\text{Subject3}}}{1} + \frac{Y_{\text{Subject4}}}{\frac{1}{2}} + \frac{Y_{\text{Subject5}}}{1} + \frac{Y_{\text{Subject6}}}{\frac{1}{2}}\right)$$

$$= \frac{1}{6}\sum_{i=1}^{6}\frac{Q_i Y_i}{P(Q_i = 1 \mid X_i)}$$

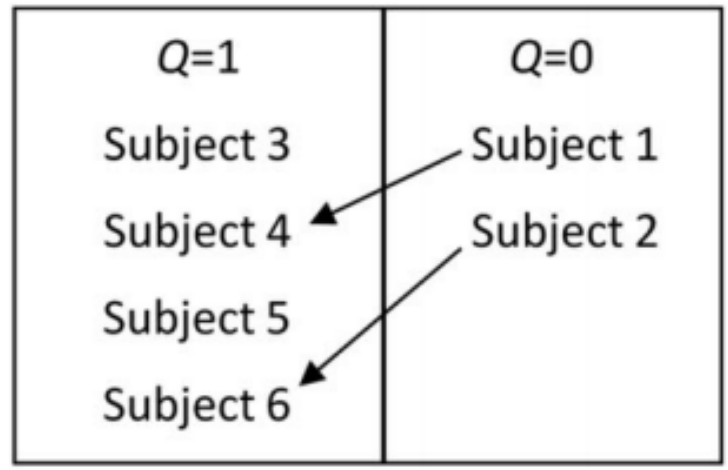


Figure 2. Grouping subjects based on having complete or missing data.

Megan M. Marron & Abdus S. Wahed (2016) Teaching Missing Data
Undergraduates Using a Group-Based Project Within a Six-Week S
Journal of Statistics Education, 24:1, 8-15

Last Observed Carried Forward (LOCF) Imputation

- Plug in last available measurement in place of the missing values
- **Advantages**: very simple
- **Disadvantages**: decreased sample variance (replacement with identical values)
- It is the least preferred method because of large bias

Subject	Age	Sex	Week		
			1	2	3
1	60	F	20.1	20.9	.
2	43	M	13.7	.	15.3
3	20	M	18.0	19.1	20.2
4	38	M	19.3	20.0	.

→

Subject	Age	Sex	Week		
			1	2	3
1	60	F	20.1	20.9	**20.9**
2	43	M	13.7	**13.7**	15.3
3	20	M	18.0	19.1	20.2
4	38	M	19.3	20.0	**20.0**

Megan M. Marron & Abdus S. Wahed (2016) Teaching Missing Data I
Undergraduates Using a Group-Based Project Within a Six-Week Su
Journal of Statistics Education, 24:1, 8-15

Unconditional Mean Imputation

- Method is to replace missing values with the mean of the available values
- **Advantages**: easy to implement
- **Disadvantages**: leads to a reduction in variability because you are imputin based on the mean. It also changes the correlation between the imputed variable vs. other variables.

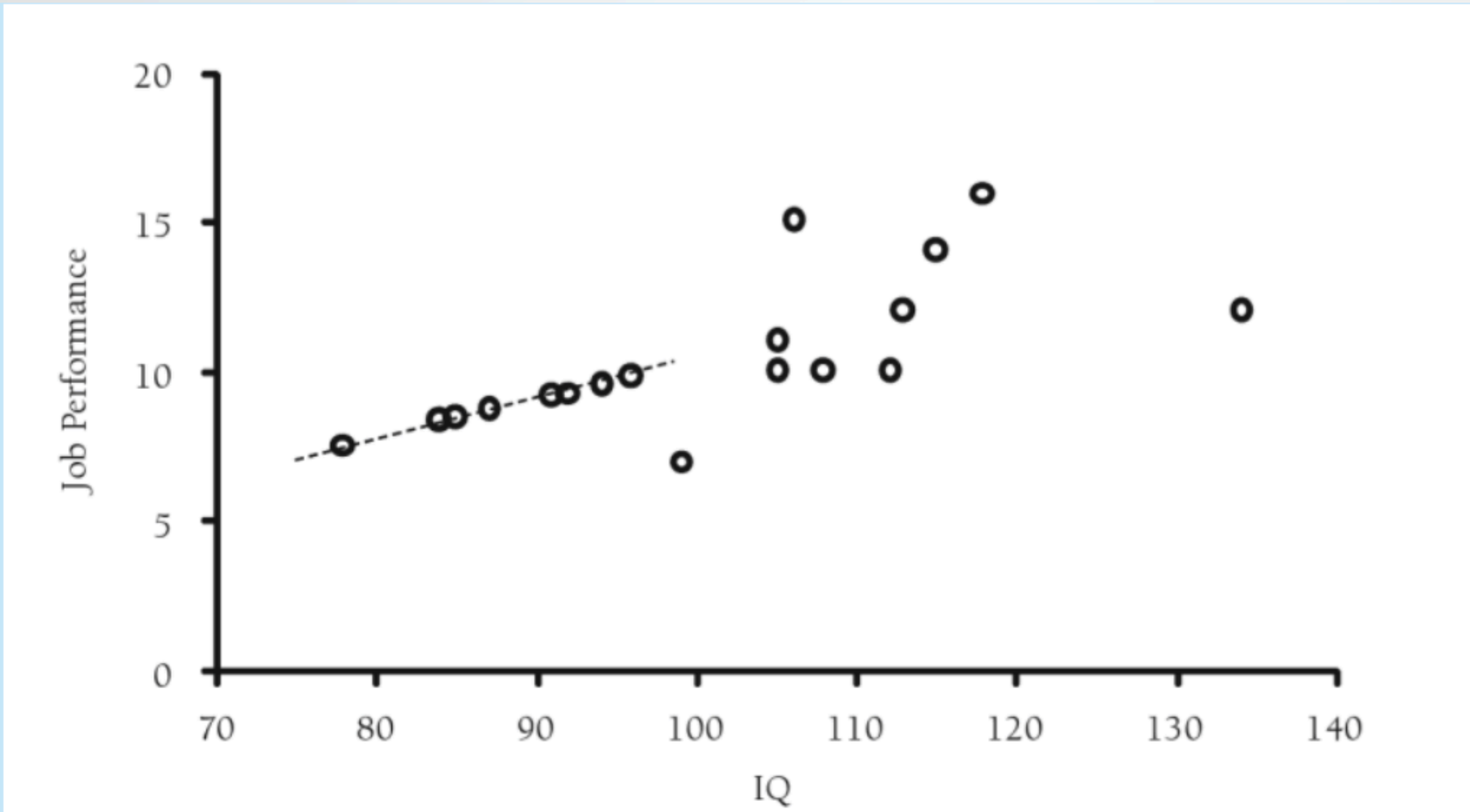
*since mean imputation is based on the values of that variable itself, there is no relationship between the imputed variable and other observed variables

Single/Deterministic/Regression Imputation

- also known as regression/conditional mean imputation: where missing values are imputed with predicted values from a regression equation
- **Advantages**: usage of complete information to impute
- **Disadvantages**: imputed values are directly from the regression line, decreasing variability. Also inflates correlations because it is using values that are already perfectly correlated with one another to impute. There is no residual variance because the imputed points fit perfectly onto the regression line. It does not refle the full uncertainty of the missing data.

*impute observed values of a variable based on values of other variables

Example of Regression Imputation

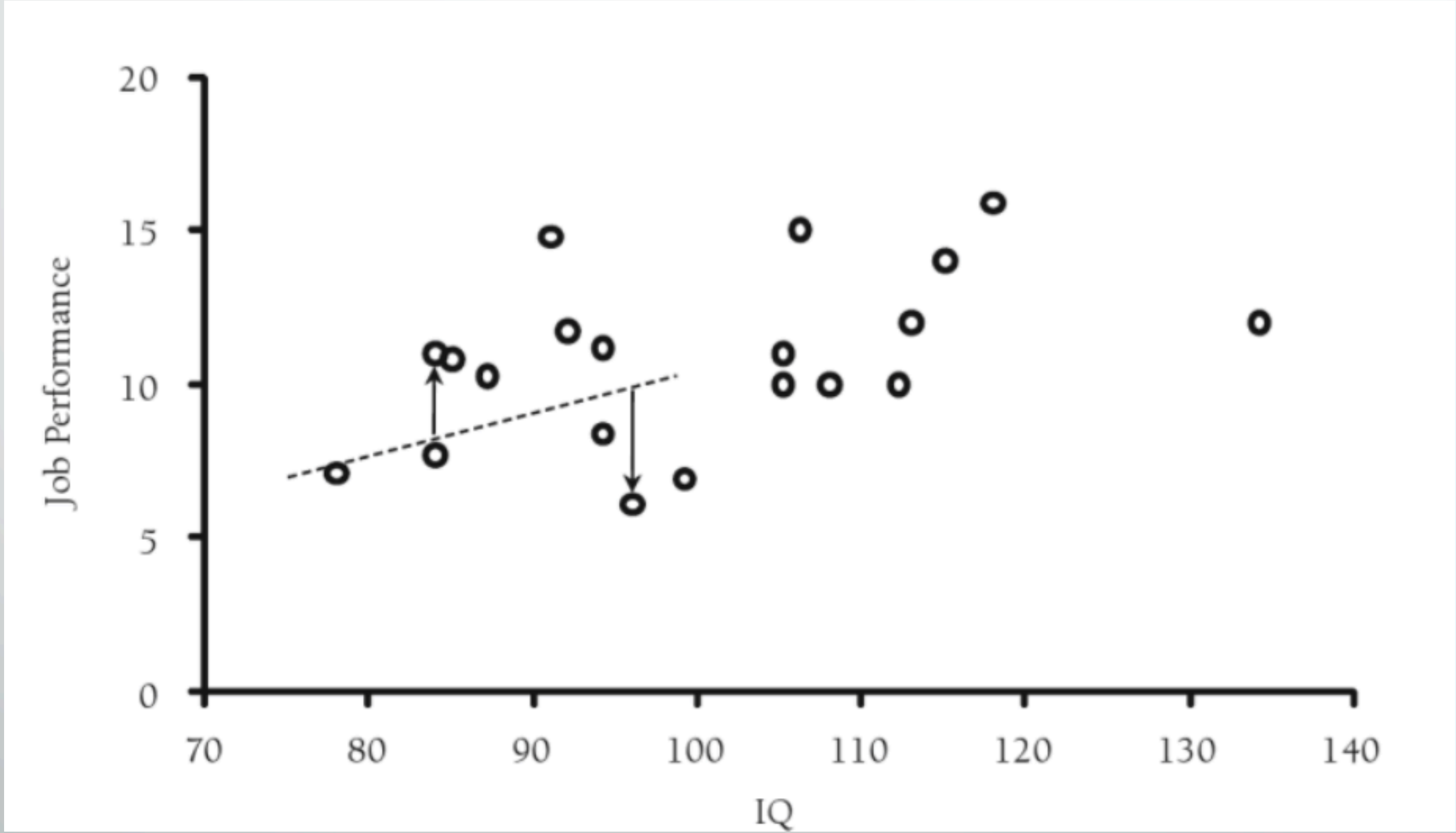


Megan M. Marron & Abdus S. Wahed (2016) Teaching Mi
Methodology to Undergraduates Using a Group-Based P
Six-Week Summer Program, Journal of Statistics Educat

Stochastic Imputation

- Addresses problems with regression imputation (lost variability) with the attemp to add back this lost variability
- Done by adding randomly drawn residuals from regression imputation, based on residual variance from regression model
- **Advantages**: "adds back" lost variability from regression imputation and produce unbiased correlation estimates under MAR
- **Disadvantages**: standard errors are less biased than in regression imputation but still weakened

Example of Stochastic Imputation



Megan M. Marron & Abdus S. Wahed (2016) Teaching Mis
Methodology to Undergraduates Using a Group-Based Pr
Week Summer Program, Journal of Statistics Education,

Multiple Imputation

- Obtain several estimates of the missing value using draws from the multivariate distribution of all variables.
- Or, at an intro level, obtain several estimates of the missing value using regression of the variable on all other variables.
- Either way, take the average of all the estimates.

Mean Imputation Example w/ R code

1. Convert Stata file of missing data into R
2. Perform mean imputation on the missing values dataset (becomes new dataset
3. Compare correlation tables of the missing values dataset and the imputed dataset
4. Compare summary statistics of the missing values dataset and the imputed dataset and observe the variability
5. Perform linear regression and observe differences

Set-up

Use the haven package to convert the Stata file into R.

```
#convert hsb2.dta and hsb2_mar.dta from Stata to R. Replicate missing dataset so we can impute.
fullData <- read_dta("~/Downloads/hsb2.dta")
View(fullData)
missingData <- read_dta("~/Downloads/hsb2_mar.dta")
View(missingData)
meanImputedData <- missingData
View(meanImputedData)
```

Preview of the full dataset:

	id	female	race	ses	schtyp type of school	prog type of program	read reading score	write writing score	math math score	science science score	socst social studies score
1	70	male	white	low	public	general	57	52	41	47	57
2	121	female	white	middle	public	vocation	68	59	53	63	61
3	86	male	white	high	public	general	44	33	54	58	31
4	141	male	white	high	public	vocation	63	44	47	53	56
5	172	male	white	middle	public	academic	47	52	57	53	61
6	113	male	white	middle	public	academic	44	52	51	63	61
7	50	male	african-amer	middle	public	general	50	59	42	53	61

Set-up (continued)

Preview of the missing values dataset:

	id	female	race	ses	schtyp type of school	prog type of program	read reading score	write writing score	math math score	science science score	socst social studies score
1	116	female	white	middle	public	NA	57	59	54	50	56
2	170	male	white	high	public	NA	47	62	61	69	66
3	97	male	white	high	public	NA	60	54	58	58	61
4	104	male	white	high	public	NA	54	63	57	55	46
5	121	female	white	middle	public	NA	68	59	53	63	61
6	94	male	white	high	public	NA	55	49	61	NA	56
7	65	female	white	middle	public	NA	55	NA	66	42	56

Preview of the imputed dataset:

	id	female	race	ses	schtyp type of school	prog type of program	read reading score	write writing score	math math score	science science score	socst social studies score
1	116	female	white	middle	public	academic	57.00000	59.00000	54.0000	50.00000	56
2	170	male	white	high	public	general	47.00000	62.00000	61.0000	69.00000	66
3	97	male	white	high	public	academic	60.00000	54.00000	58.0000	58.00000	61
4	104	male	white	high	public	vocation	54.00000	63.00000	57.0000	55.00000	46
5	121	female	white	middle	public	general	68.00000	59.00000	53.0000	63.00000	61
6	94	male	white	high	public	academic	55.00000	49.00000	61.0000	51.30978	56
7	65	female	white	middle	public	academic	55.00000	52.95082	66.0000	42.00000	56

Perform Mean Imputation

- For each column (variable), use a for loop and iterate through each observation. If an observation matches as missing (NA), we set it equal to the mean of the whole column.
- Example:

```
#for read variable
for(i in meanImputedData$id){
  if(is.na(meanImputedData[i,7]) == TRUE){
    meanImputedData[i,7] = mean(meanImputedData$read, na.rm = TRUE)
  }
}
#for math variable
for(i in meanImputedData$id){
  if(is.na(meanImputedData[i,9]) == TRUE){
    meanImputedData[i,9] = mean(meanImputedData$math, na.rm = TRUE)
  }
}
```

- Repeat for each variable with missing values

Analyze Correlation Tables

- Using the cor() function and selecting the columns that we want to analyze, we compare the corr
tables for the missing values dataset and the imputed dataset. We can observe altered correlation

Correlation table for full dataset:

	read	write	math	science	socst
read	1.0000000	0.5967765	0.6622801	0.6301579	0.6214843
write	0.5967765	1.0000000	0.6174493	0.5704416	0.6047932
math	0.6622801	0.6174493	1.0000000	0.6307332	0.5444803
science	0.6301579	0.5704416	0.6307332	1.0000000	0.4651060
socst	0.6214843	0.6047932	0.5444803	0.4651060	1.0000000

Correlation table for missing values dat

	read	write	math	science
read	1.0000000	0.5807117	0.6478505	0.6232614
write	0.5807117	1.0000000	0.6175467	0.5734457
math	0.6478505	0.6175467	1.0000000	0.6409446
science	0.6232614	0.5734457	0.6409446	1.0000000
socst	0.5740294	0.5744264	0.5200243	0.4218290

Correlation table for the imputed dataset:

	read	write	math	science	socst
read	1.0000000	0.5480112	0.6158825	0.6076032	0.6028521
write	0.5480112	1.0000000	0.5491474	0.4955650	0.5707113
math	0.6158825	0.5491474	1.0000000	0.5576771	0.5106852
science	0.6076032	0.4955650	0.5576771	1.0000000	0.4306481
socst	0.6028521	0.5707113	0.5106852	0.4306481	1.0000000

Summary Statistics for Mean Imputation

Here, we compare the summary statistics with a focus on mean and standard deviation

For full dataset:

	read	write	math	science	socst
mean	52.23000	52.775000	52.645000	51.850000	52.40500
sd	10.25294	9.478586	9.368448	9.900891	10.73579
min	28.00000	31.000000	33.000000	26.000000	26.00000
max	76.00000	67.000000	75.000000	74.000000	71.00000

For missing values dataset:

	read	write	math	science
mean	52.28796	52.950820	52.897297	51.309783
sd	10.21072	9.257773	9.360837	9.817833
min	28.00000	31.000000	33.000000	26.000000
max	76.00000	67.000000	75.000000	74.000000

For mean imputed data:

	read	write	math	science	socst
mean	52.28796	52.950820	52.89730	51.309783	52.40500
sd	9.97715	8.853514	9.00113	9.414877	10.73579
min	28.00000	31.000000	33.00000	26.000000	26.00000
max	76.00000	67.000000	75.00000	74.000000	71.00000

Question: What do we notice about the summary statistics for the mean imputed data when cor to the other datasets? Specifically the standard deviation?

Summary Statistics for Mean Imputation

Here, we compare the summary statistics with a focus on mean and standard deviation

For full dataset:

	read	write	math	science	socst
mean	52.23000	52.775000	52.645000	51.850000	52.40500
sd	10.25294	9.478586	9.368448	9.900891	10.73579
min	28.00000	31.000000	33.000000	26.000000	26.00000
max	76.00000	67.000000	75.000000	74.000000	71.00000

For missing values dataset:

	read	write	math	science
mean	52.28796	52.950820	52.897297	51.309783
sd	10.21072	9.257773	9.360837	9.817833
min	28.00000	31.000000	33.000000	26.000000
max	76.00000	67.000000	75.000000	74.000000

For mean imputed data:

	read	write	math	science	socst
mean	52.28796	52.950820	52.89730	51.309783	52.40500
sd	9.97715	8.853514	9.00113	9.414877	10.73579
min	28.00000	31.000000	33.00000	26.000000	26.00000
max	76.00000	67.000000	75.00000	74.000000	71.00000

Answer: We see that the mean stays the same because we imputed based on the mean of the o values. More interestingly, we see that the standard deviation is less for each variable in the imp This is because we imputed based on the center of the distribution, decreasing the variability.

Linear Regression for Mean Imputation

- Perform bivariate regression analysis using the lm() and summary() functions

Graph of the relationship between math and write (dependent) with regression lines for all 3 datasets

Full data= RED
Missing values data = BLUE
Mean Imputed values data =

Regression Results for Imputation Methods

	Full Data	CCA	Mean Imputation	Mean Imputation w/ Cat.	Multiple Imputation
Intercept	9.62**	13.03**	13.94***	9.11*	10.11**
write	0.37***	0.44***	0.38***	0.33***	0.38***
math	0.44***	0.32***	0.34***	0.46***	0.42***
female	-2.70*	-2.71*	-2.17(.)	-0.59	-2.67*
prog general	0.23	0.52	1.72	1.51	0.63
prog academic	1.88	1.81	2.95(.)	2.35(.)	2.42
R-squared	0.5045	0.4679	0.4257	0.4449	0.5147
# of missing observations	0	70	35	0	0

Standard Deviations/Means/Proportions for Imputation Meth

	Full Data	CCA	Mean Imputation	Mean Imputation w/ Cat.	Multiple Imputat
read	mean: 52.23 sd: 10.253	mean: 52.288 sd: 10.211	mean: 52.288 sd: 9.977	mean: 52.288 sd: 9.977	mean: sd:
write	mean: 52.775 sd: 9.479	mean: 52.951 sd: 9.258	mean: 52.951 sd: 8.854	mean: 52.951 sd: 8.854	mean: sd:
math	mean: 52.645 sd: 9.368	mean: 52.897 sd: 9.361	mean: 52.897 sd: 9.001	mean: 52.897 sd: 9.001	mean: sd:
science	mean: 51.85 sd: 9.901	mean: 51.310 sd: 9.818	mean: 51.310 sd: 9.415	mean: 51.310 sd: 9.415	mean: sd:
female	male: 91/200 = 0.455 female: 109/200 = 0.545	male: 81/182 = 0.445 female: 101/182 = 0.555	male: 81/182 = 0.445 female: 101/182 = 0.555	male: 91/200 = 0.455 female: 109/200 = 0.545	male: female:
prog	vocation: 50/200 = 0.25 general: 45/200 = 0.225 academic: 105/200 = 0.525	vocation: 46/182 = 0.253 general: 41/182 = 0.225 academic: 95/182 = 0.522	vocation: 46/182 = 0.253 general: 41/182 = 0.225 academic: 95/182 = 0.522	vocation: 50/200 = 0.25 general: 42/200 = 0.21 academic: 108/200 = 0.54	vocation: general: academic:
# of missing observations	0	70	35	0	0

***Notice again how the standard deviations for the variables in the mean imputed dataset are**

Standard Errors of Coefficients

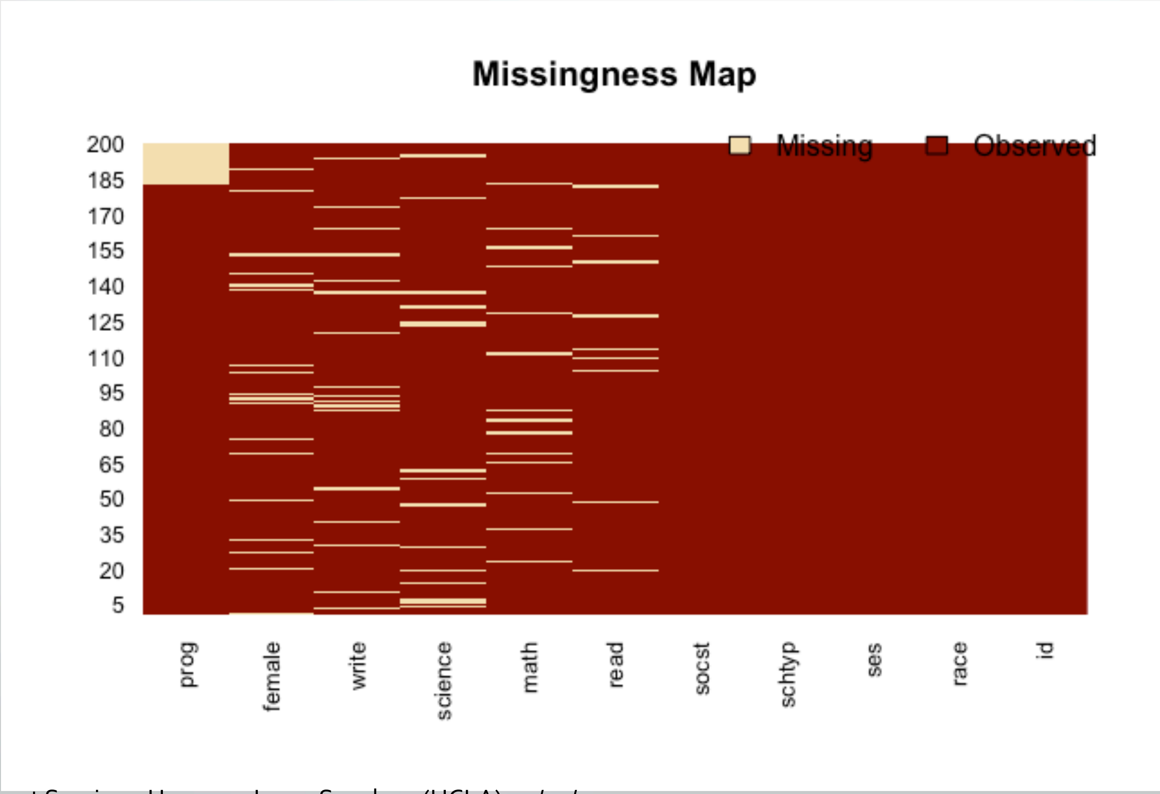
	Full Data	CCA	Mean Imputation	Mean Imputation w/ Cat.	Multiple Imputati
Intercept	3.40980	4.12355	3.94658	3.78154	3.49099
write	0.07463	0.09265	0.08193	0.07492	0.08237
math	0.07500	0.09514	0.07986	0.07503	0.08270
female	1.09541	1.36519	1.18717	1.10094	1.23014
prog general	1.51219	1.88083	1.66568	1.53230	1.62628
prog academic	1.42307	1.65486	1.50343	1.41087	1.52243

Visualization in R

- Packages that can be used to visualize the missing data through plots include VIM and Amelia
- VIM
  - spineMiss
  - aggr
  - matrixplot
  - marginplot
- Amelia
  - missmap

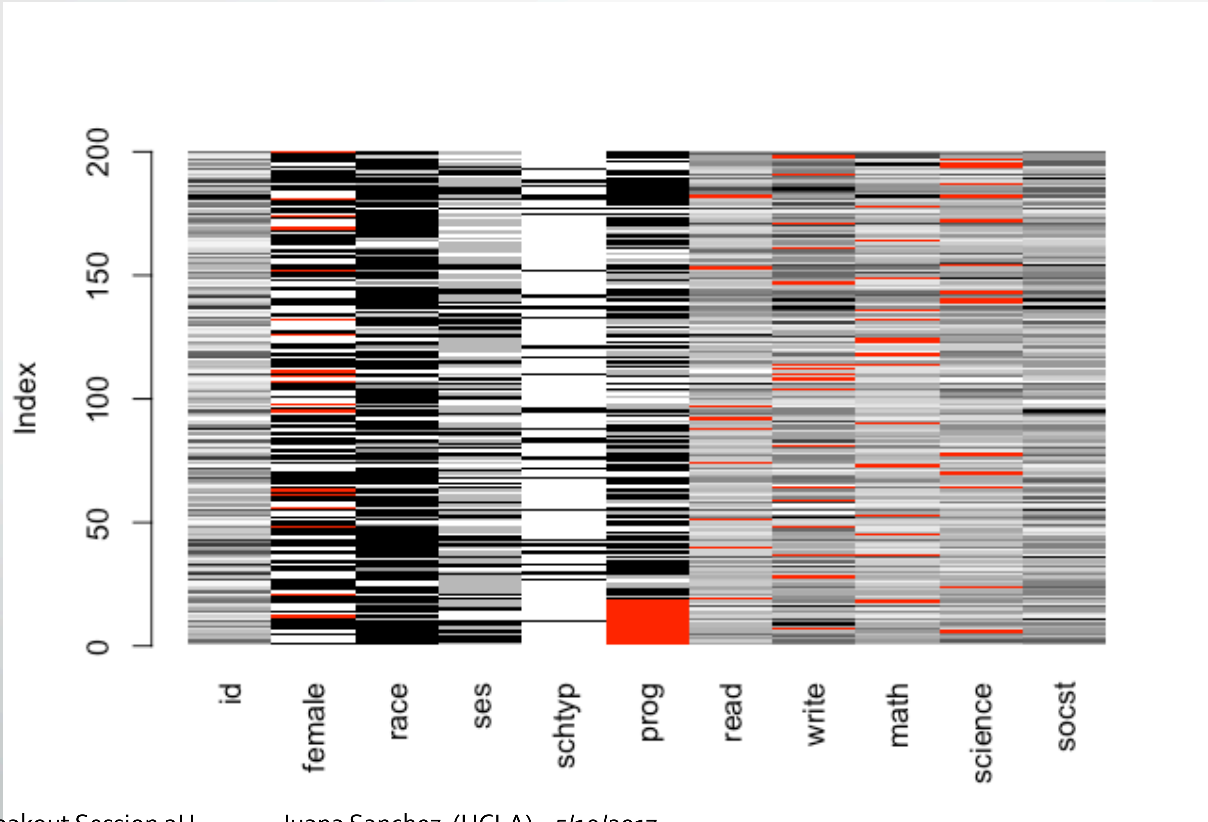
missmap

missmap → creates a simple plot showing where the missing data occurs in the dataset. Can be used to observe patterns



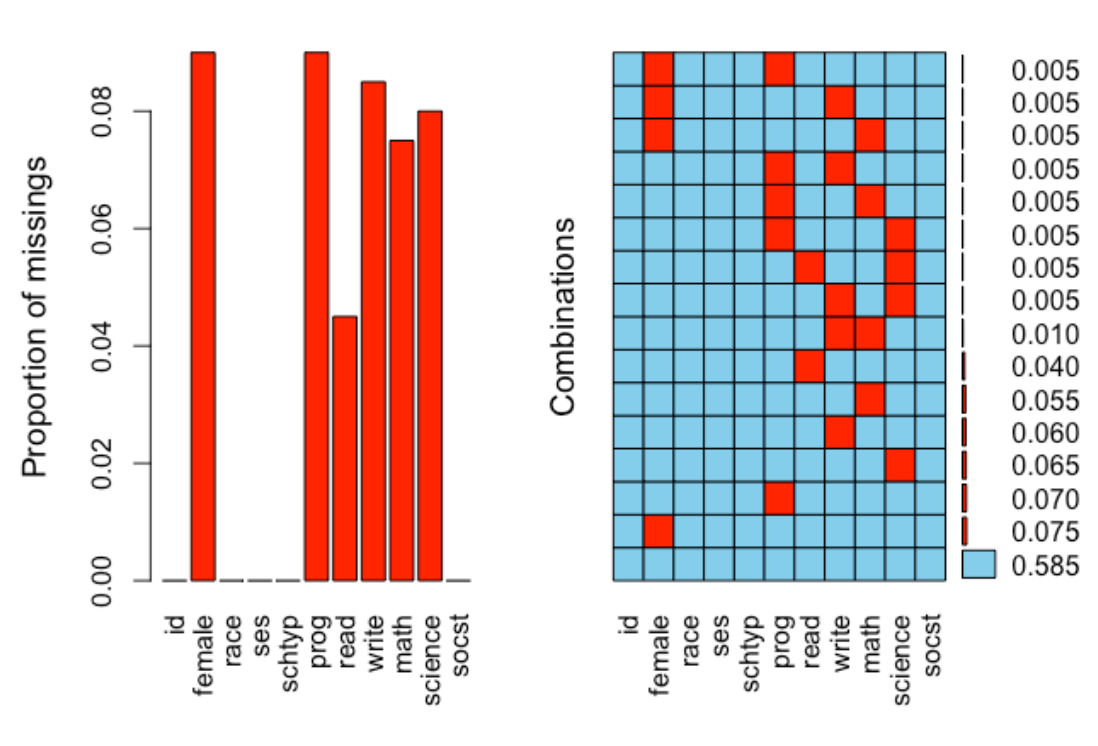
matrixplot

- matrixplot → available data is coded according to a continuous color scheme, while missi data is visualized with a distinguishable color (red)



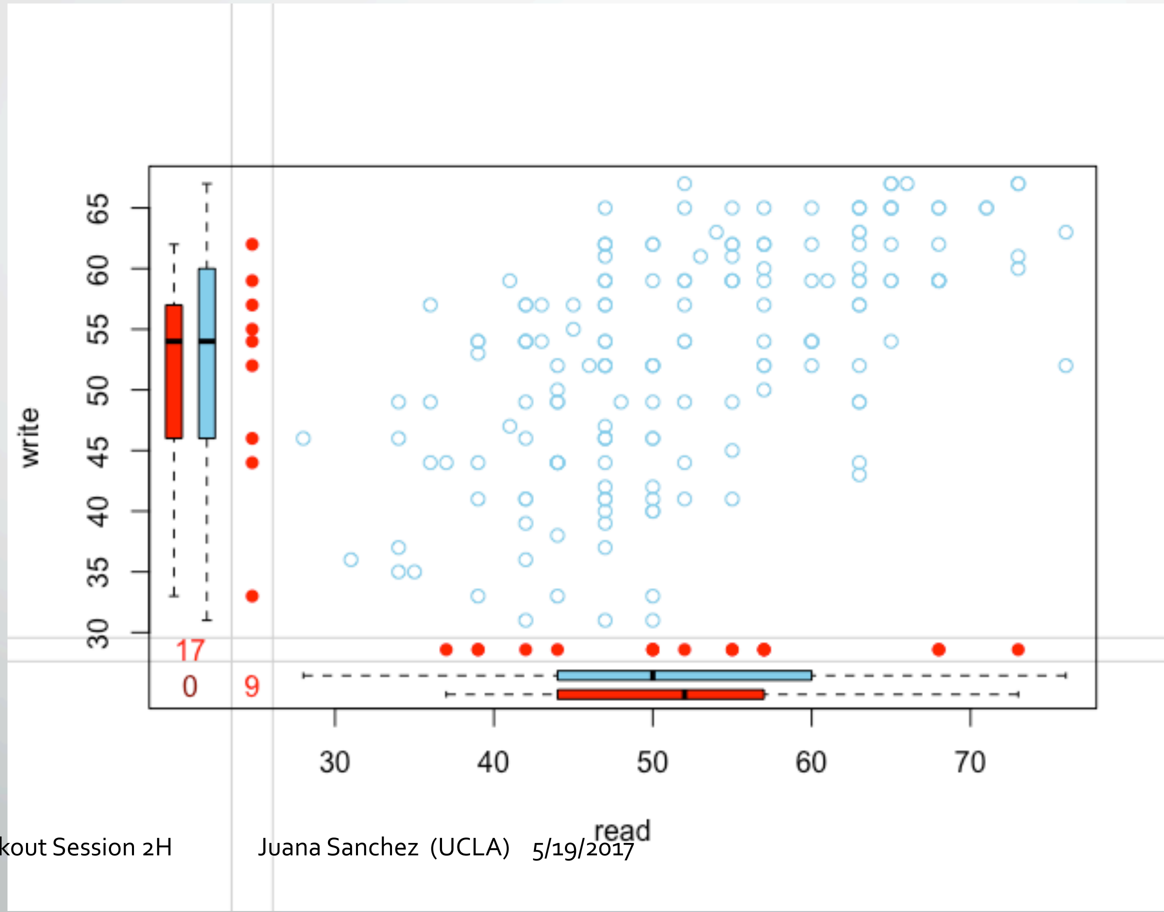
aggr

- aggr → shows patterns of missingness through a combinations graph, with red representing missing data. For example, 0.585 of the observations are not missing any and 0.075 are missing just female.



marginplot

- marginplot → enhanced scatterplot which also shows boxplots for each variable, observ and imputed



"Well, no, I don't see any patterns in this data, but I did see Elvis in my oatmeal this morning!"

https://www.cartoonstock.com/directory/b/busi

Multiple Imputation in Cancer Research

- Citation: Royston P (2004). "Multiple Imputation of Missing Values." The Stata Journal, 4,
  227–241.
- Dataset: http://www.stata-press.com/data/r13/brcancer.dta
- Research Interest: Recurrence-free survival time (duration in years from entry into study t
  time of death or disease recurrence)
- Total observations: 686
- Author created a new dataset from brcancer.dta (full) called brcaex.dta with missing valu
  (20% of observations were completely at random missing)
- Standard errors were larger with imputed data, and parameter estimates are similar

Multiple Imputation in Sociology Research

- Citation: Finke R, Adamczyk A (2008). "Cross-National Moral Beliefs: The Influence of National Religious Context." Sociological Quarterly, 49(4), 617–652.
- Dataset: One from ISSP (https://dbk.gesis.org/dbksearch/sdesc2.asp?no=3190) and one from WVS (can't access)
- Research Interest: How does religion relate to morality on a micro to macro level? (state vs. national)
- For the ISSP dataset, there were 39034 observations with 35356 after excluding ones with missing info on key variables
- "Approximately twenty percent of respondents in each dataset were missing information on variables needed in the analysis. In a preliminary analysis we ran all models with listwise deleted data, pairwise deleted data, and multiply imputed data, and found that our results were minimally affected by these different techniques for handling missing da Since multiple imputation takes full advantage of the available data and avoids some of the bias in standard errors and test statistics that can accompany pairwise deletion, we chose to present our results using multiple imputation"

Multiple Imputation in Occupational Health

- Citation: Chamberlain LJ, Crowley M, Tope D, Hodson R (2008). "Sexual Harassment in Organizational Context." Work and Occupations, 35(3), 262–295.
- Dataset: available on http://www.icpsr.umich.edu/icpsrweb/ICPSR/studies/3979
- 204 observations; 51 with missing variables
- Research Interest: How does workplace dignity affect employee work behaviors and organizational performance?
- Did not explicitly state how results changed after MICE

Multiple Imputation in Obesity/Physical Activity Research

- Citation: Wiles NJ, Jones GT, Haase AM, Lawlor DA, Macfarlane GJ, Lewis G (2008). "Physical Activity and Emotional Problems Amongst Adolescents." Social Psychiatry and Psychiatric Epidemiology, 43(10), 765–772.
- Dataset: ?
- Research Interest: Relationship between physical activity and emotional problems in children aged 11-14 in England
- Total observations: 1424 children, with 206 missing at 1 year follow-up
- " Imputing missing data using MICE suggested that those imputed were more likely to have higher scores at follow-up. Sensitivity analyses including imputed data were consistent with th results of the complete-case analyses suggesting that missing data had not biased the results"

Multiple Imputation in Behavior Research

- Citation: Melhem NM, Brent DA, Ziegler M, Iyengar S, Kolko D, Oquendo M, Birmaher B, Burke A, Zelazny Stanley B, Mann, J J (2007). "Familial Pathways to Early-Onset Suicidal Behavior: Familial and Individual Antecedents of Suicidal Behavior." American Journal of Psychiatry, 164(9), 1364–1370.
- Dataset: ?
- Research Interest: identify clinical predictors of new-onset suicidal behavior in children of parents with a history of mood disorder and suicidal behavior
- 17% of the sample had no missing data for any variable, 31% had missing data for one or two variables, 43 had missing data for three or four variables, and 9% had missing data for more than four variables
- Results?

Multiple Imputation in Health Economics Research

- Citation: Burton A, Billingham LJ, Bryan S (2007). "Cost-Effectiveness in Clinical Trials: Using Multiple Imputation to Deal with Incomplete Cost Data." Clinical Trials, 4(2), 154–161.
- Dataset: ?
- Research Interest: The objective of this article is to investigate the appropriateness and practicalities of using MI to handle missing cost component data as an alternative to the standard complete case analysis, when one of the aims of the trial is to assess cost effectivene
- 115 observations, 82 with complete data
- The complete case analysis resulted in a higher mean cost for those patients randomized to M + PC of £2804 (95% CI £1236 to £4290) compared to PC. When MI was used, a smaller differen between treatments in terms of the mean cost of £2384 was seen (95% CI £833 to £3954).

Summary

- There are several imputation methods to replace missing data with substituted data, as well as several mechanisms that govern which imputation methods we should use
- We investigated CCA (complete case analysis), mean imputation, and multiple imputation in depth and how each method affected results
- We can visualize missing data using packages in R that produce visually appealing plots and graphs

Citations

- Megan M. Marron & Abdus S. Wahed (2016) Teaching Missing Data Methodology to Undergraduates Using a Group-Based Project Within a Six-Week Summer Program, Journal of Statistics Education, 24:1, 8-15
- http://www.ats.ucla.edu/stat/stata/seminars/missing_data/Multiple_imputation/mi_in_stata_pt1_new.htm