Missing Data

November 25, 2019

Dana Tomc Dougherty

Agenda

01
Motivation

Education (in years)	Income (in $1000s)
20	71.2
13	NaN
14	NaN
18	NaN
16	41.8
...	
15	51.6

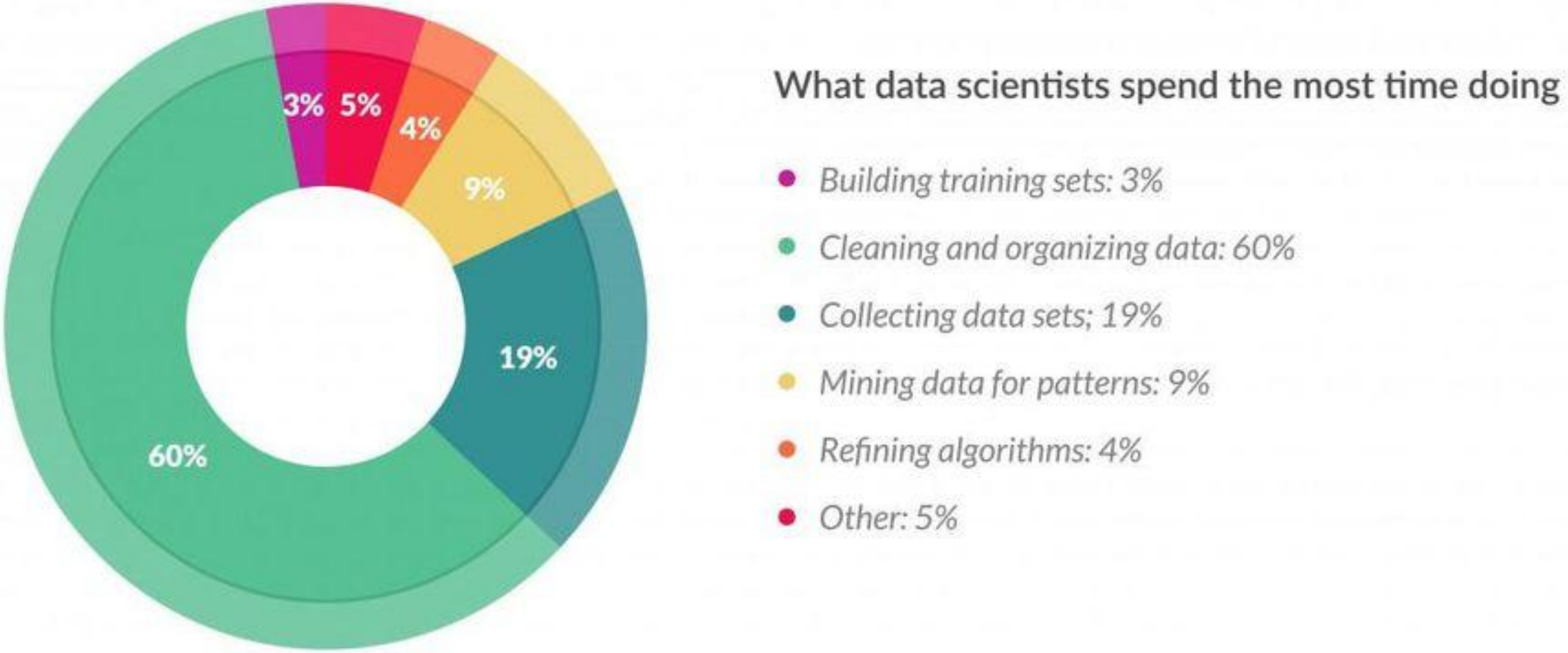
Motivation

- "Garbage in, garbage out"
- Selection bias → Biased results
- Loss of information → Biased results

  → Loss of statistical power

Education (in years)	Income (in $1000s)
20	71.2
13	NaN
14	NaN
18	NaN
16	41.8
...	
15	51.6

Motivation

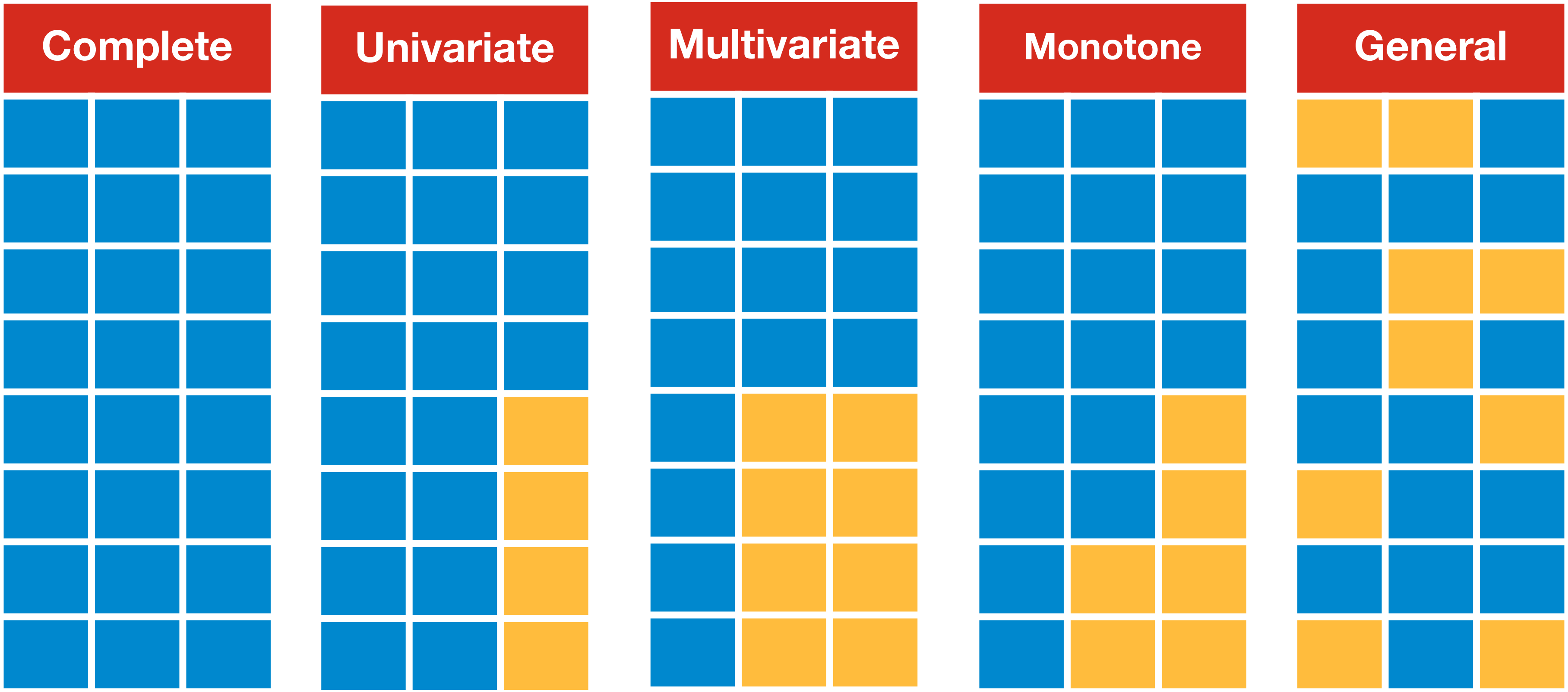


Source: forbes.com

02
Background

Background

Missing Data Patterns



Background
Missingness Mechanisms

Missing Completely at Random (MCAR)

Missing at Random (MAR)

Missing Not at Random (MNAR)

Background

Missingness Mechanisms

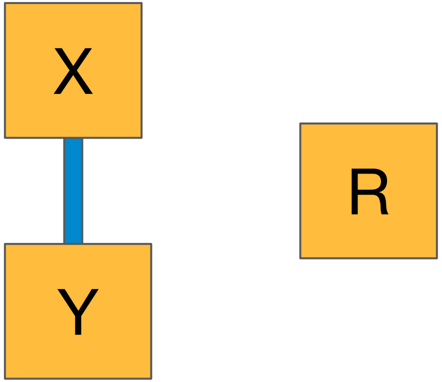
X = observed part of the data

Y = missing part of the data

R = missing data indicator

X = Education (in years)	Y = Income (in $1000s)	R
20	71.2	0
13	NaN	1
14	NaN	1
18	NaN	1
16	41.8	0
	...	
15	51.6	0

Background
Missingness Mechanisms



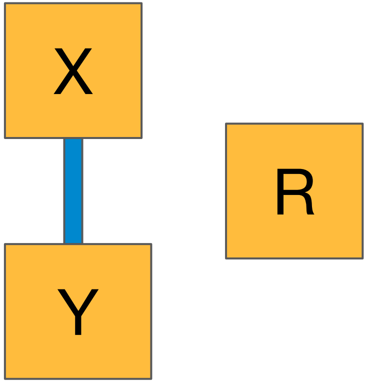
X = Education (in years)	Y = Income (in $1000s)	R
20	71.2	0
13	NaN	1
14	NaN	1
18	NaN	1
16	41.8	0
	...	
15	51.6	0

Background

Missingness Mechanisms

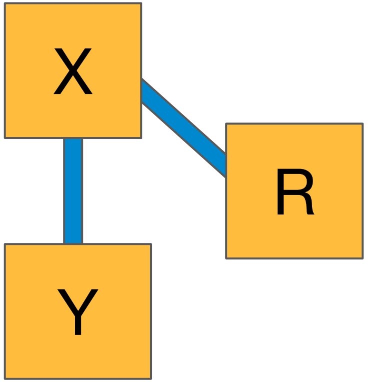
Missing Completely at Random (MCAR)

$P(R \mid X, Y) = P(R)$



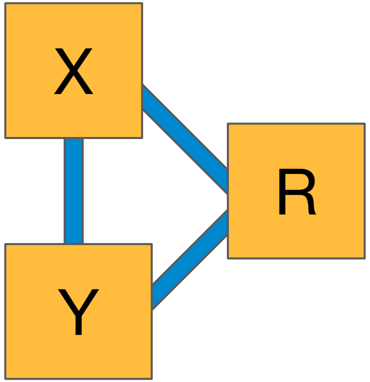
Missing at Random (MAR)

$P(R \mid X, Y) = P(R \mid X)$



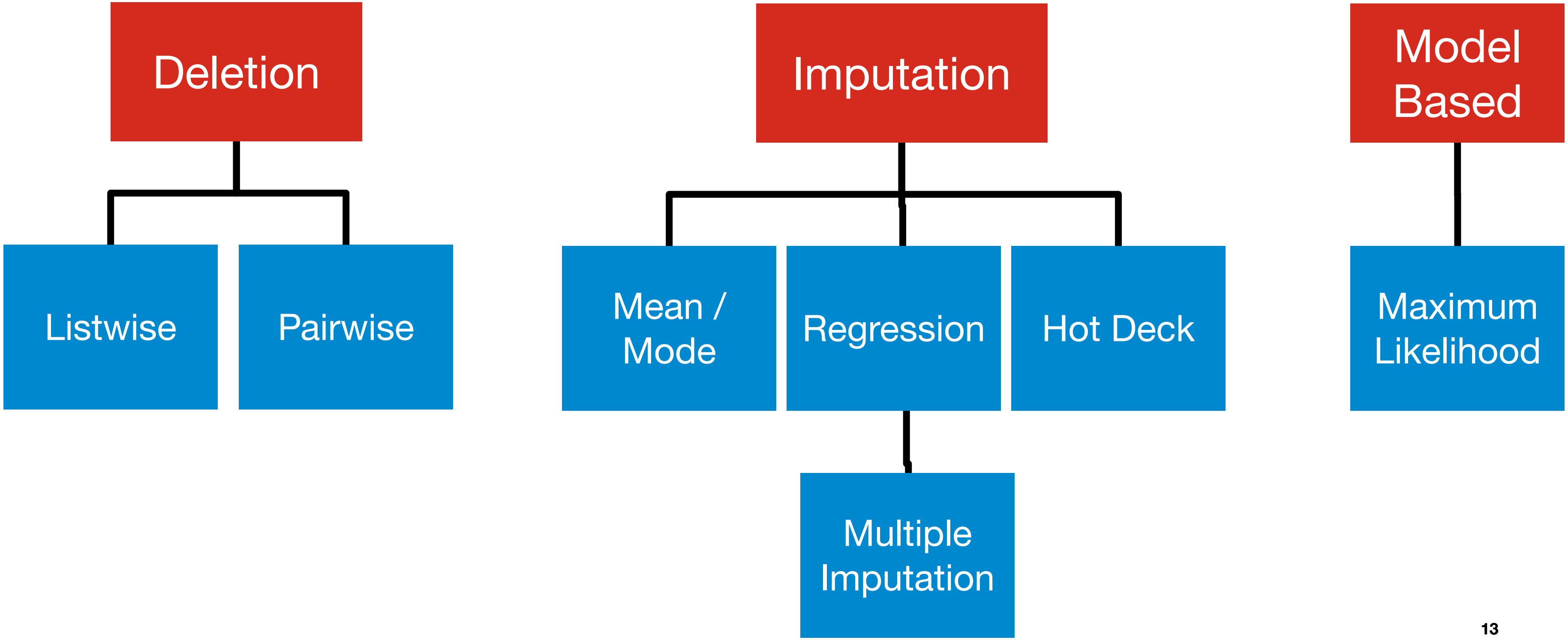
Missing Not at Random (MNAR)

$P(R \mid X, Y) = P(R \mid X, Y)$



Background

Missing Data Techniques



03
Simulation

Simulation

Create a complete dataset → Apply missingness mechanism → Apply missing data technique → Analyze data →

Simulation

Create a complete dataset → **Apply missingness mechanism** → **Apply missing data technique** → **Analyze data** →

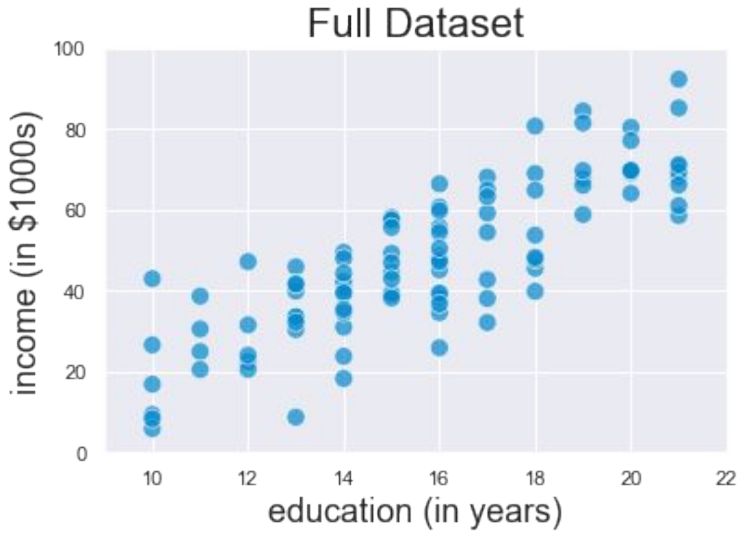
Education (in years)	Income (in $1000s)
20	71.2
13	30.2
14	33.1
18	60.1
16	41.8
...	
15	51.6

- N rows
- *income* ~ Normal(48, 400)
- *education* = 8 + 0.17 * *income* + *e*
  e ~ Normal(0, 4)
  discretized education

Simulation

Create a complete dataset → **Apply missingness mechanism** → **Apply missing data technique** → **Analyze data** →

Education (in years)	Income (in $1000s)
20	71.2
13	30.2
14	33.1
18	60.1
16	41.8
...	
15	51.6



Simulation

Create a complete dataset → **Apply missingness mechanism** → **Apply missing data technique** → **Analyze data** →

Remove 25% of the data

1. MCAR
2. MAR
3. MNAR

Simulation

Create a complete dataset → **Apply missingness mechanism** → **Apply missing data technique** → **Analyze data** →

1. **MCAR**
   - Randomly select n rows
   - Delete *income* from those rows

Education (in years)	Income (in $1000s)
20	71.2
13	30.2
14	33.1
18	60.1
16	41.8
15	51.6

→

Education (in years)	Income (in $1000s)
20	NaN
13	NaN
14	33.1
18	60.1
16	NaN
15	51.6

Simulation

Create a complete dataset → **Apply missingness mechanism** → Apply missing data technique → Analyze data →

2. MAR

- Select n rows where *education* is the largest
- Delete *income* from those rows

Education (in years)	Income (in $1000s)
20	71.2
13	30.2
14	33.1
18	60.1
16	41.8
15	51.6

Education (in years)	Income (in $1000s)
20	NaN
13	30.2
14	33.1
18	NaN
16	NaN
15	51.6

Simulation

Create a complete dataset → **Apply missingness mechanism** → **Apply missing data technique** → **Analyze data**

3. MNAR

- Select n rows where *income* is the largest
- Delete *income* from those rows

Education (in years)	Income (in $1000s)
20	71.2
13	30.2
14	33.1
18	60.1
16	41.8
15	51.6

→

Education (in years)	Income (in $1000s)
20	NaN
13	30.2
14	33.1
18	NaN
16	41.8
15	NaN

Simulation

Create a complete dataset → **Apply missingness mechanism** → **Apply missing data technique** → **Analyze data** →

1. Listwise Deletion
2. Mean Imputation
3. Regression Imputation
4. Multiple Imputation
5. Maximum Likelihood

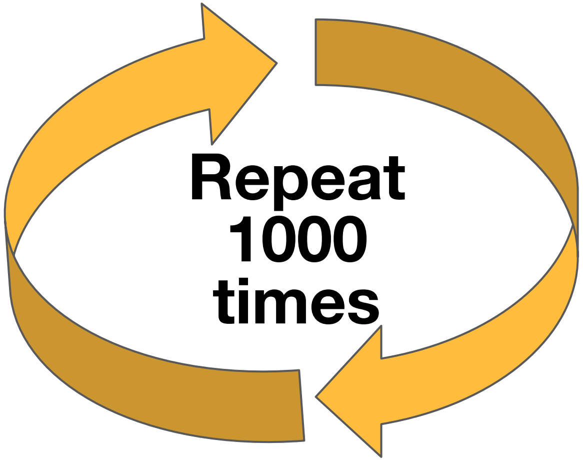
Simulation

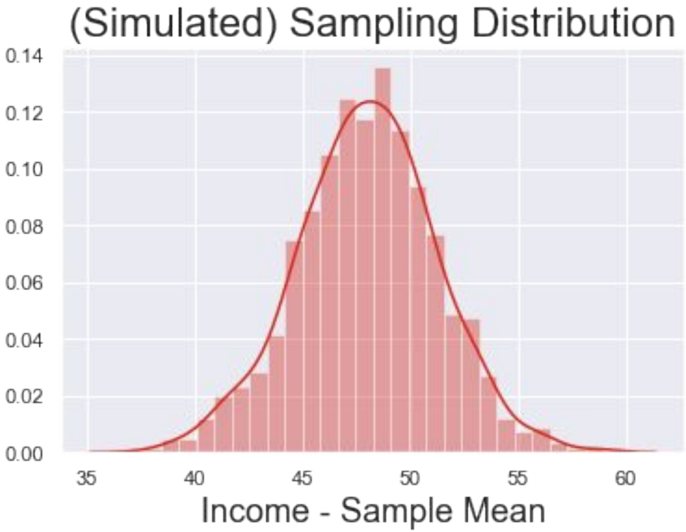
Create a complete dataset → **Apply missingness mechanism** → **Apply missing data technique** → **Analyze data** →

Calculate the sample mean of *income*

Simulation

Create a complete dataset → Apply missingness mechanism → Apply missing data technique → Analyze data →



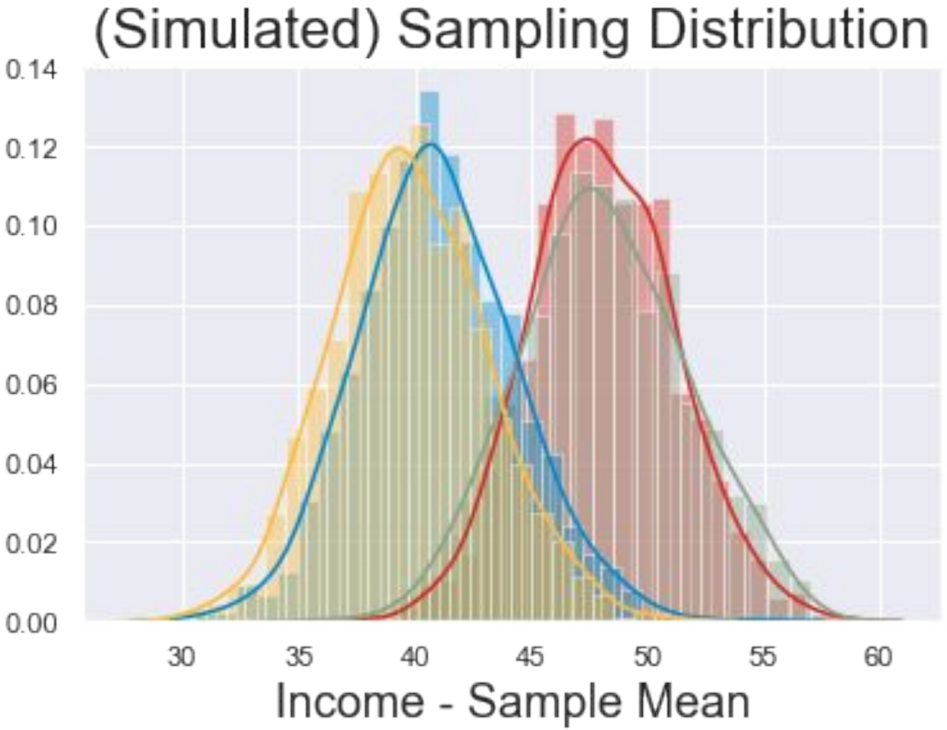


$$E(\overline{Y}) = \mu = 48$$

$$\sigma^2(\overline{Y}) = \frac{\sigma^2}{n} = \frac{20^2}{40} = 10$$

Simulation

Sampling Distribution of (Income) Sample Mean

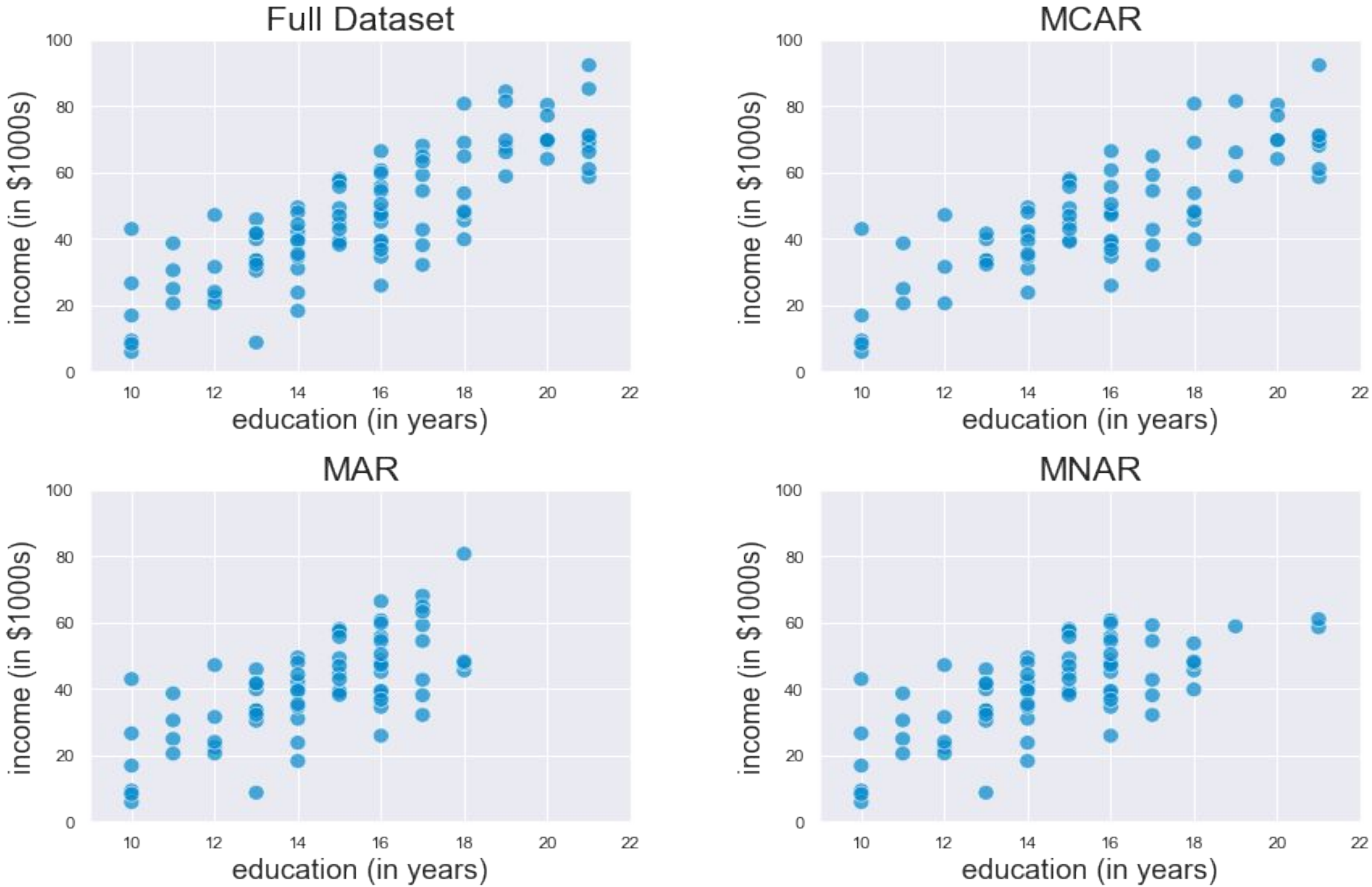


Listwise Deletion

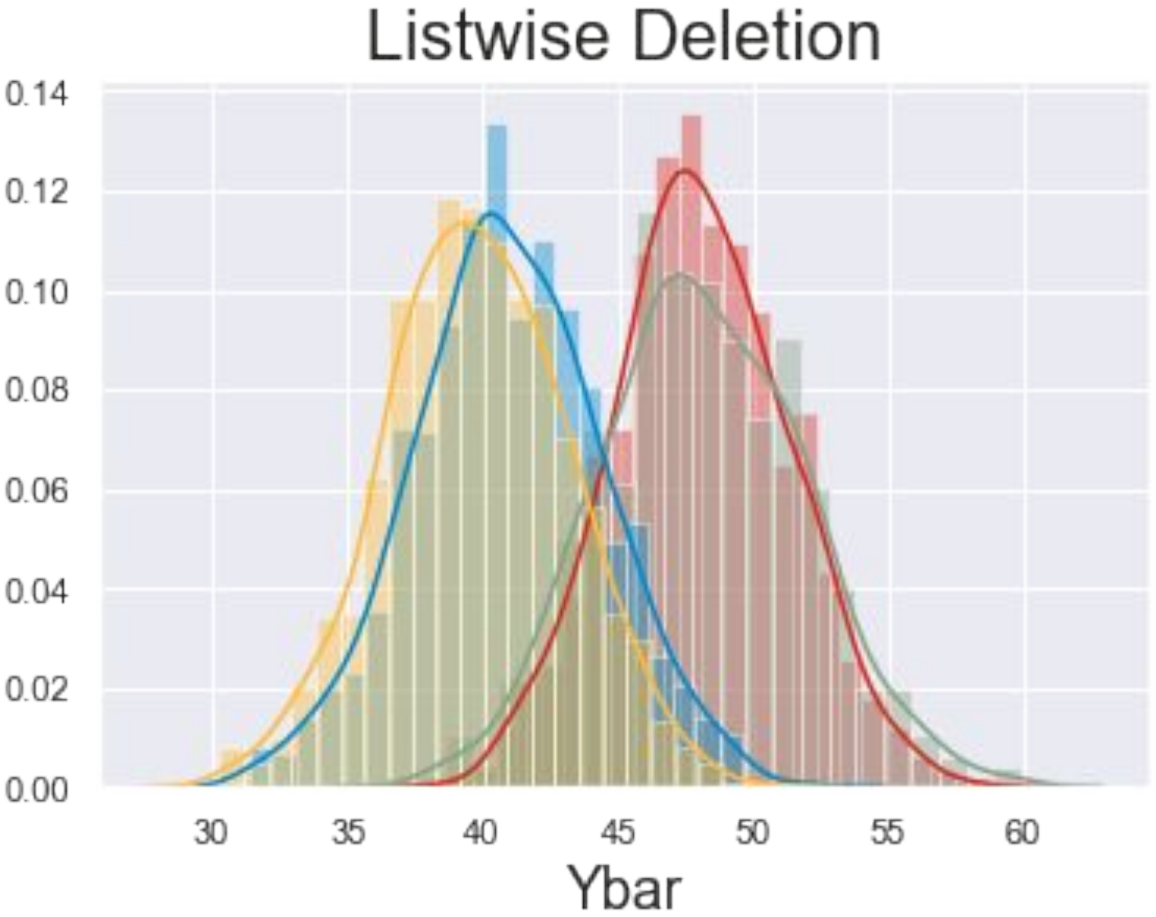
Listwise Deletion

Education (in years)	Income (in $1000s)
20	71.2
13	30.2
14	33.1
18	NaN
16	NaN
15	NaN

Education (in years)	Income (in $1000s)
20	71.2
13	30.2
14	33.1



Full Dataset
MCAR
MAR
MNAR
income (in $1000s)
education (in years)
0
20
40
60
80
100
10
12
14
16
18
20
22



Dataset	Mean	Variance
Full	48.07	10.07
MCAR	48.11	14.06
MAR	40.93	11.87
MNAR	39.78	11.07

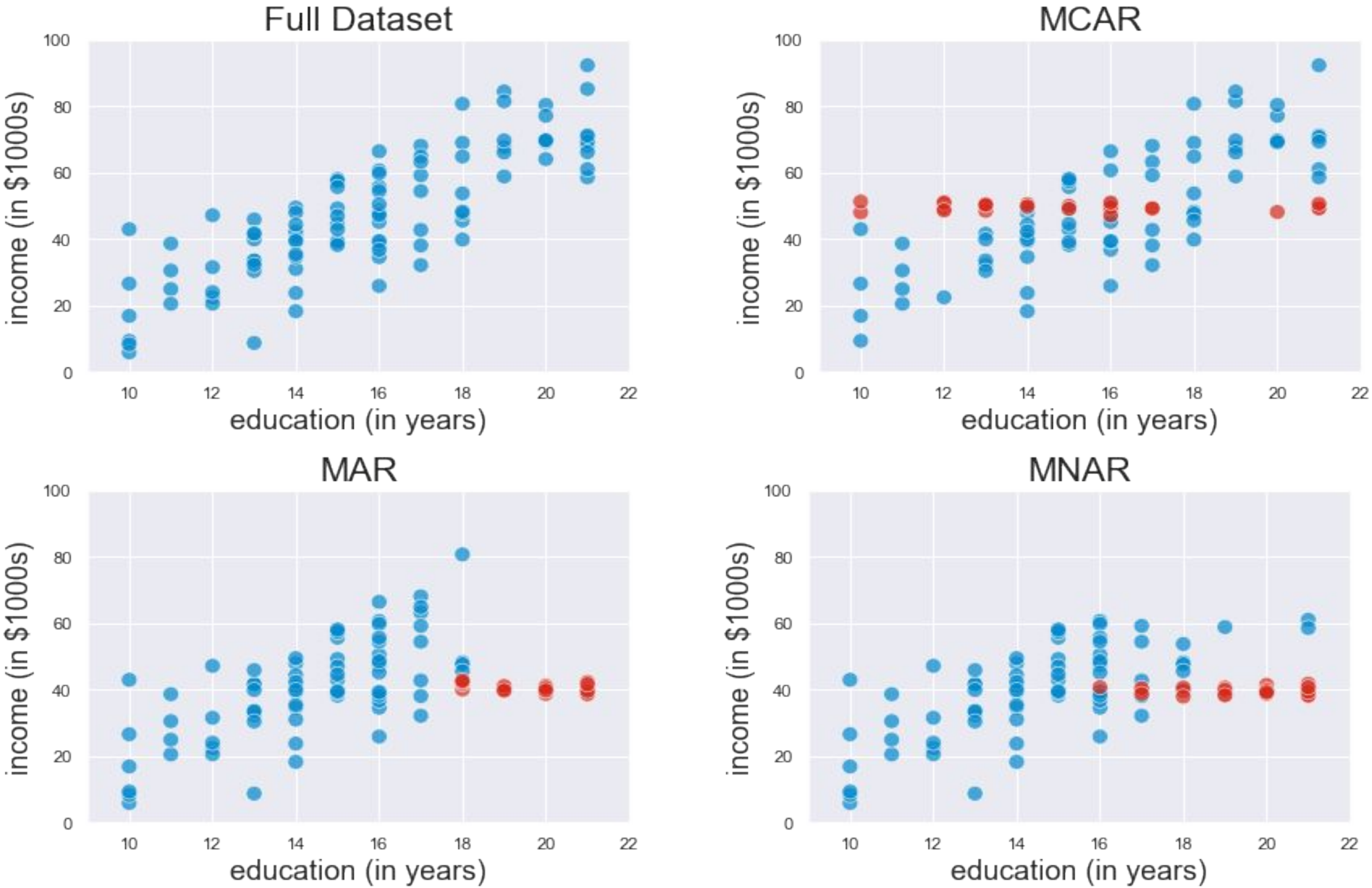
Mean Imputation

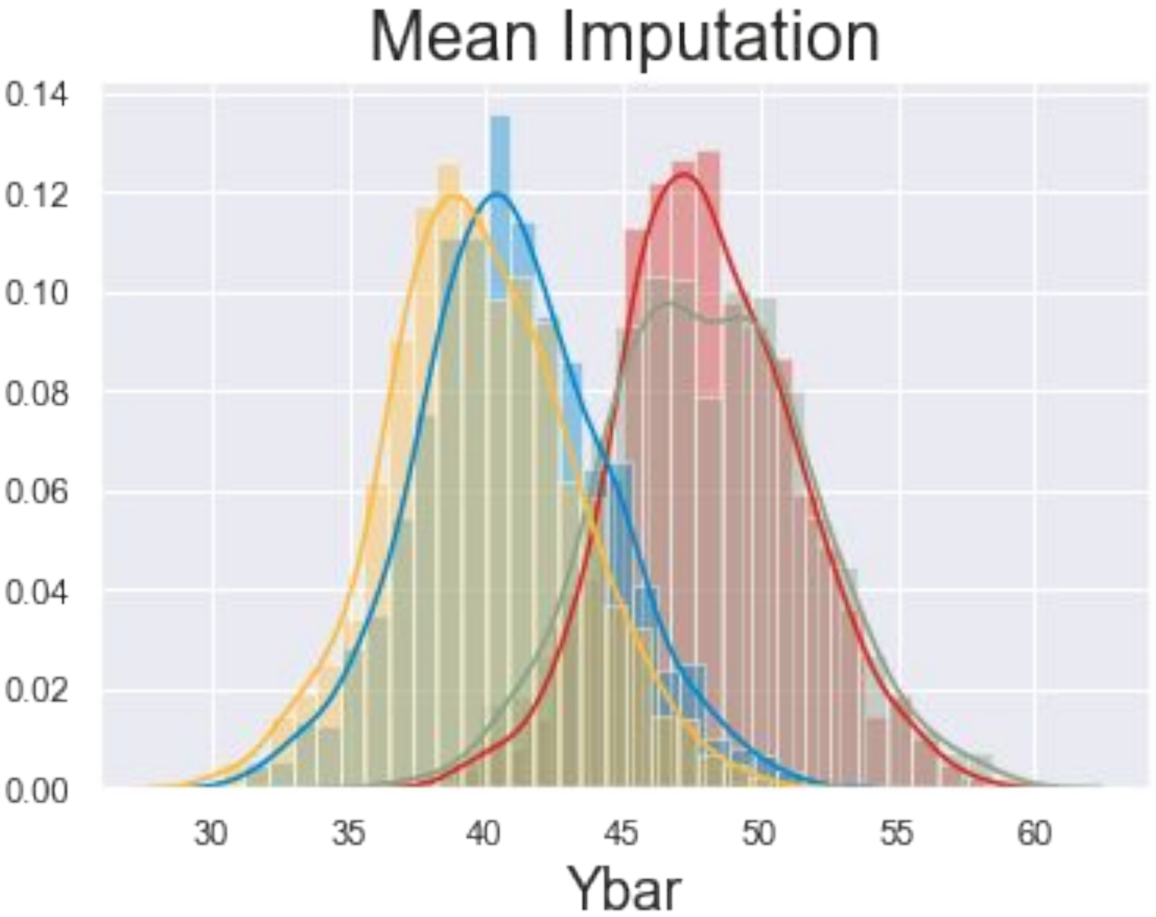
Mean Imputation

Education (in years)	Income (in $1000s)
20	71.2
13	30.2
14	33.1
18	NaN
16	NaN
15	NaN

Education (in years)	Income (in $1000s)
20	71.2
13	30.2
14	33.1
18	44.8
16	44.8
15	44.8

$$\frac{71.2 + 30.2 + 33.1}{3} = 44.8$$


Full Dataset
MCAR
MAR
MNAR
income (in $1000s)
education (in years)
0
20
40
60
80
100
10
12
14
16
18
20
22



Dataset	Mean	Variance
Full	48.07	10.51
MCAR	48.08	13.91
MAR	40.94	11.84
MNAR	39.76	11.36

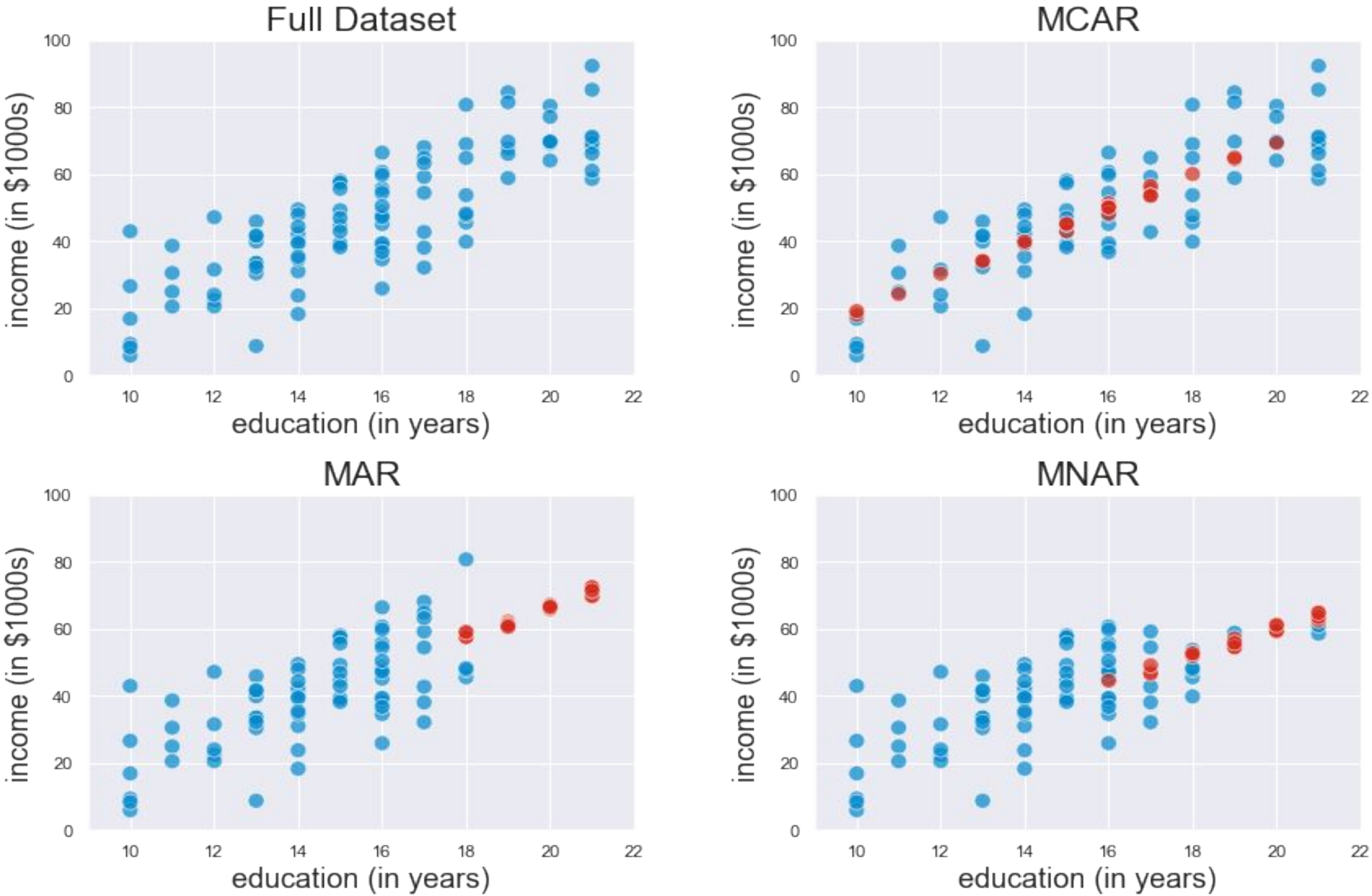
Regression Imputation

Regression Imputation

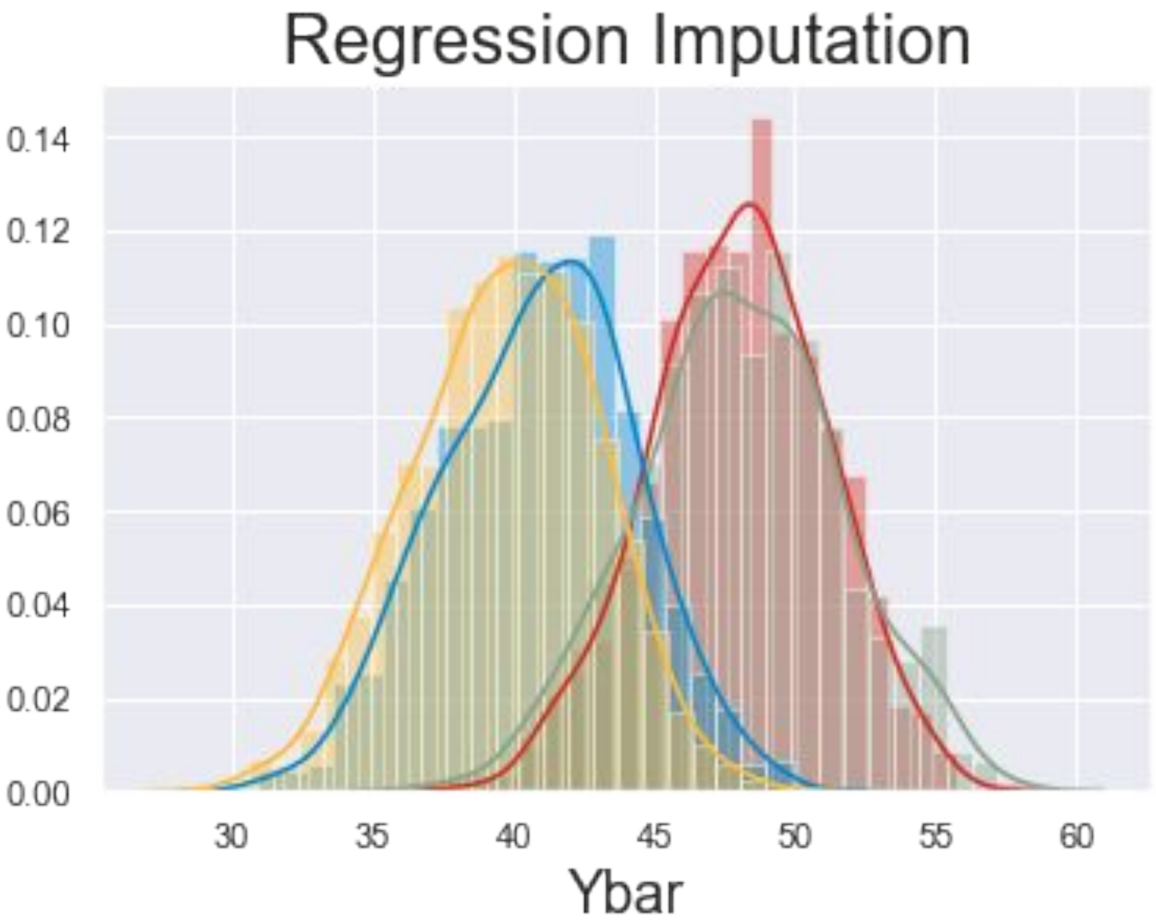
Education (in years)	Income (in $1000s)
20	71.2
13	30.2
14	33.1
18	NaN
16	NaN
15	NaN

Education (in years)	Income (in $1000s)
20	71.2
13	30.2
14	33.1
18	58.7
16	46.7
15	40.7

$$\widehat{income} = -49.6 + 6.0 * education$$



Full Dataset
MCAR
MAR
MNAR
income (in $1000s)
education (in years)



Dataset	Mean	Variance
Full	47.99	9.49
MCAR	48.04	13.03
MAR	40.89	11.44
MNAR	39.71	10.73

Multiple Imputation

Multiple Imputation

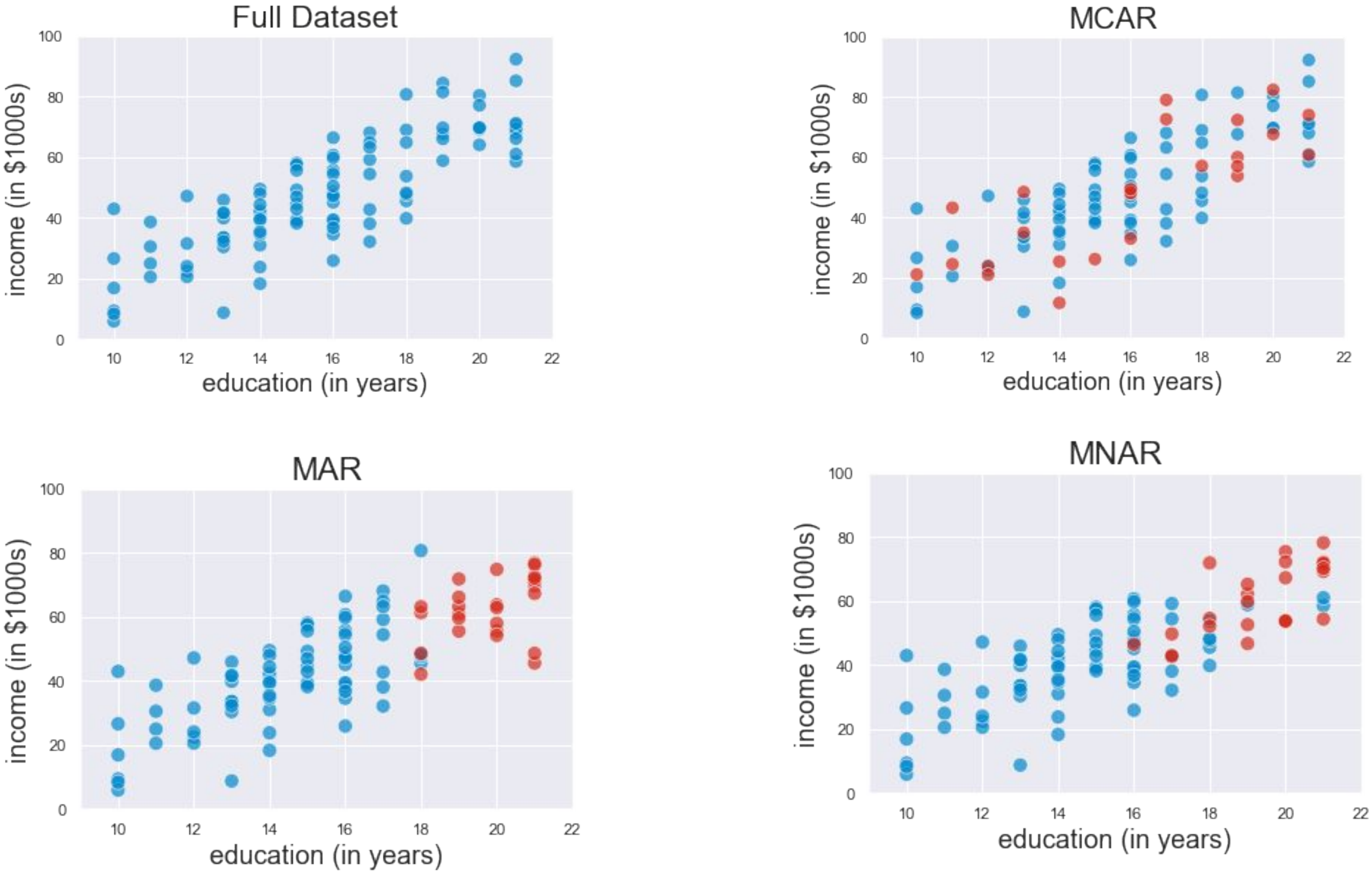
Education (in years)	Income (in $1000s)
20	71.2
13	30.2
14	33.1
18	NaN
16	NaN
15	NaN

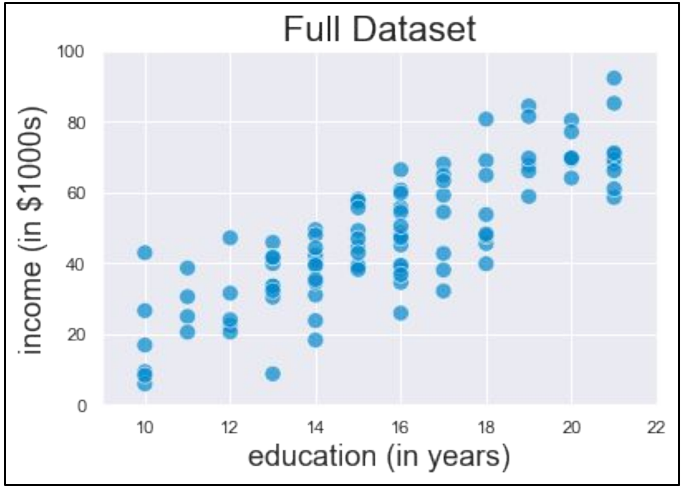
Education (in years)	Income (in $1000s)
20	71.2
13	30.2
14	33.1
18	59.7
16	47.2
15	41.4

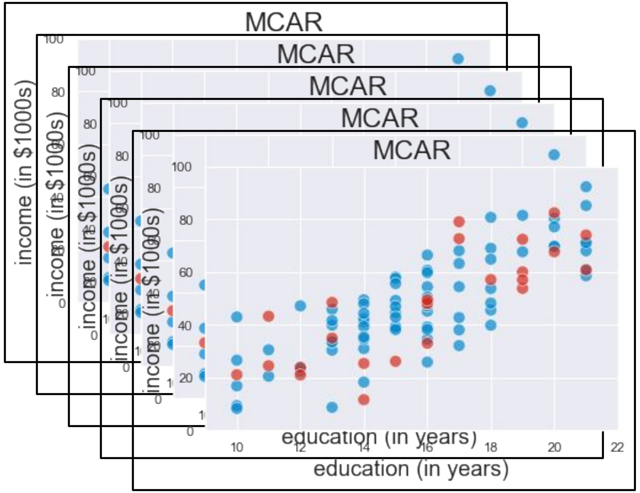
$$\overline{income}_1 = 49.1$$
$$\overline{income}_2 = 45.6$$
$$\overline{income}_3 = 47.9$$
$$\overline{income}_4 = 51.2$$
$$\overline{income}_5 = 47.1$$

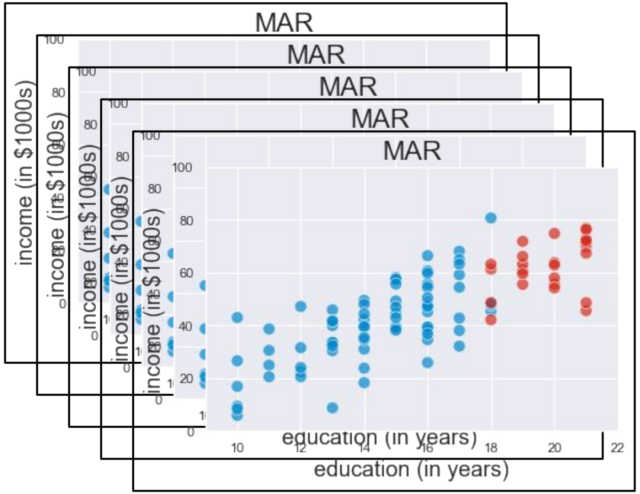
$$\overline{income}_{total} = 48.2$$

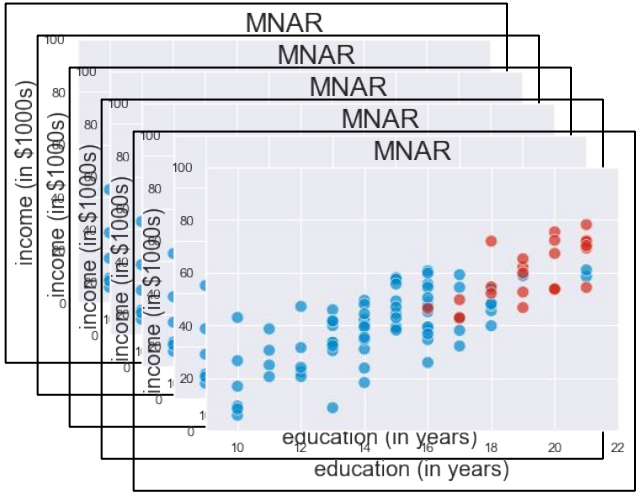
$$\widehat{income} = -49.6 + 6.0 * education + e$$

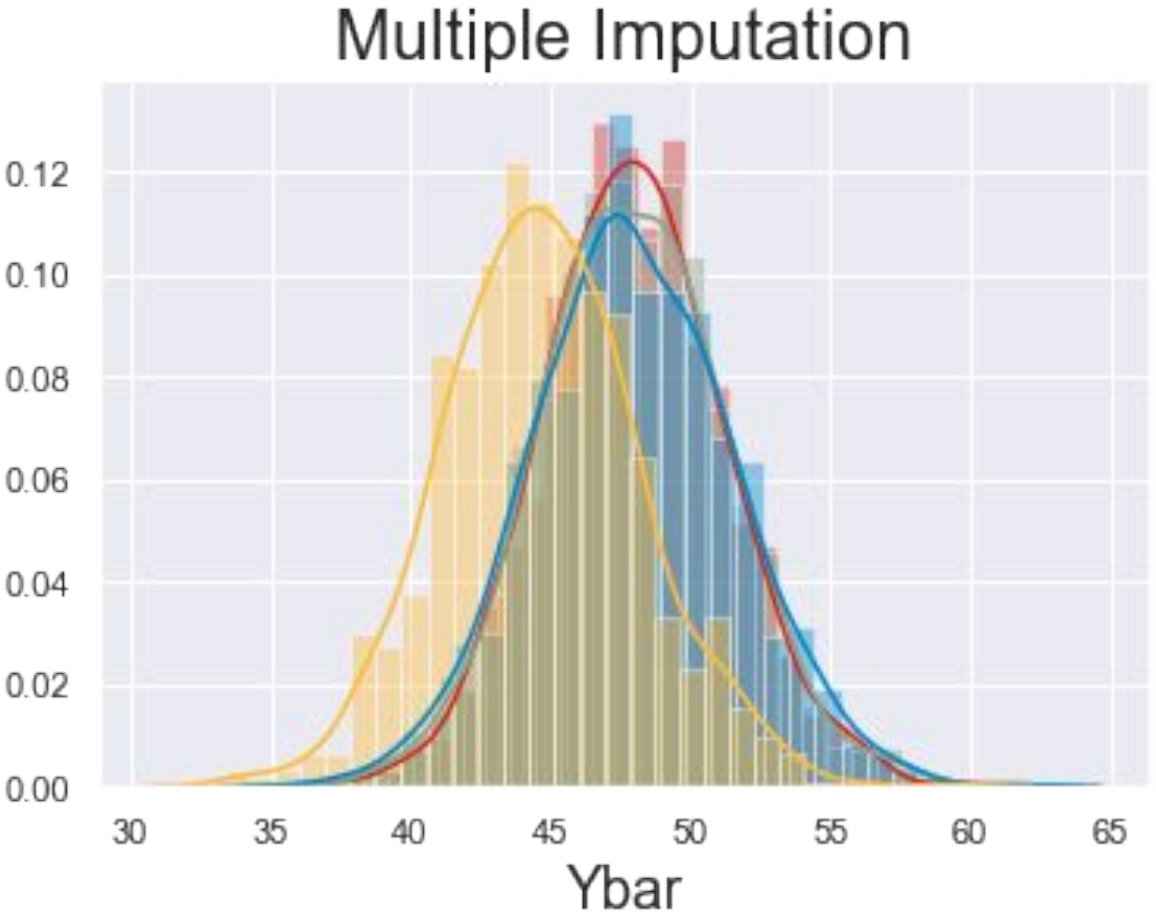

Full Dataset
MCAR
MAR
MNAR
income (in $1000s)
education (in years)
0
20
40
60
80
100
10
12
14
16
18
20
22


Full Dataset
income (in $1000s)
education (in years)
0
20
40
60
80
100
10
12
14
16
18
20
22


MCAR
income (in $1000s)
education (in years)


MAR
income (in $1000s)
education (in years)


MNAR
income (in $1000s)
education (in years)



Dataset	Mean	Variance
Full	47.98	10.6
MCAR	48.00	11.59
MAR	47.98	12.87
MNAR	44.68	12.39

Maximum Likelihood Estimation

Maximum Likelihood Estimation
Background

- Recall:

  income ~ Normal(48, 400)

  education = 8 + 0.17 * *income* + *e*
  e ~ Normal(0, 4)

- Normal distribution:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

Education (in years)	**Income (in $1000s)**
20	71.2
13	30.2
14	33.1
18	60.1
16	41.8
15	51.6

Maximum Likelihood Estimation

Background

- Likelihood Function:

$$L = \prod_{i=1}^{n} \left[\frac{1}{\sigma\sqrt{2\pi}} \, exp\left(- \frac{(x_i - \mu)^2}{2\sigma^2} \right) \right]$$

- Maximum Likelihood Estimates:

$$\hat{\mu}_Y = \frac{\Sigma Y}{N} \qquad \hat{\sigma}_Y^2 = \frac{\Sigma(Y - \hat{\mu}_Y)^2}{N}$$

Maximum Likelihood Estimation

Background

Education (in years)	Income (in $1000s)
20	71.2
13	30.2
14	33.1
18	60.1
16	41.8
15	51.6

$$\hat{\mu} = \begin{pmatrix} 16.0 \\ 48.0 \end{pmatrix}$$

$$\hat{\Sigma} = \begin{pmatrix} 5.7 & 32.8 \\ 32.8 & 212.5 \end{pmatrix}$$

EM Algorithm
Background

X -Education (in years)	Y - Income (in $1000s)
20	71.2
13	30.2
14	33.1
18	NaN
16	NaN
15	NaN

EM Algorithm

$$\hat{\mu} = \begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}$$

$$\hat{\Sigma} = \begin{pmatrix} \sigma^2_X & \sigma_{X,Y} \\ \sigma_{Y,X} & \sigma^2_Y \end{pmatrix}$$

EM Algorithm

Background

1 Initialize $\hat{\mu}$ and $\hat{\Sigma}$

2 E-Step: Estimate missing values

3 M-Step: Maximize (update) $\hat{\mu}$ and $\hat{\Sigma}$

Repeat until $\hat{\mu}$ and $\hat{\Sigma}$ converges

EM Algorithm
Background

1 Initialize $\hat{\mu}$ and $\hat{\Sigma}$

2 E-Step: Estimate missing values

3 M-Step: Maximize (update) $\hat{\mu}$ and $\hat{\Sigma}$

X -Education (in years)	Y - Income (in $1000s)
20	71.2
13	30.2
14	33.1
18	NaN
16	NaN
15	NaN

$$\hat{\mu} = \begin{pmatrix} 16.0 \\ 44.8 \end{pmatrix} \qquad \hat{\Sigma} = \begin{pmatrix} 14.3 & 0 \\ 0 & 523.5 \end{pmatrix}$$

EM Algorithm

Background

1 Initialize $\hat{\mu}$ and $\hat{\Sigma}$

2 E-Step: Estimate missing values

3 M-Step: Maximize (update) $\hat{\mu}$ and $\hat{\Sigma}$

$$\hat{\beta}_1 = \frac{\hat{\sigma}_{X,Y}}{\hat{\sigma}_X^2}$$

$$\hat{\beta}_0 = \hat{\mu}_Y - \hat{\beta}_1 \hat{\mu}_X$$

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

X -Education (in years)	Y - Income (in $1000s)
20	71.2
13	30.2
14	33.1
18	59.2
16	45.8
15	42.1

EM Algorithm

Background

1 **Initialize** $\hat{\mu}$ **and** $\hat{\Sigma}$

2 **E-Step: Estimate missing values**

3 **M-Step: Maximize (update)** $\hat{\mu}$ **and** $\hat{\Sigma}$

X -Education (in years)	Y - Income (in $1000s)
20	71.2
13	30.2
14	33.1
18	59.2
16	45.8
15	42.1

Maximum Likelihood Estimates:

$$\hat{\mu}_Y = \frac{\Sigma Y}{N}$$

$$\hat{\sigma}_Y^2 = \frac{1}{N}\left(\Sigma Y^2 - \frac{(\Sigma Y)^2}{N}\right)$$

$$\hat{\sigma}_{X,Y} = \frac{1}{N}\left(\Sigma XY - \frac{\Sigma X \Sigma Y}{N}\right)$$

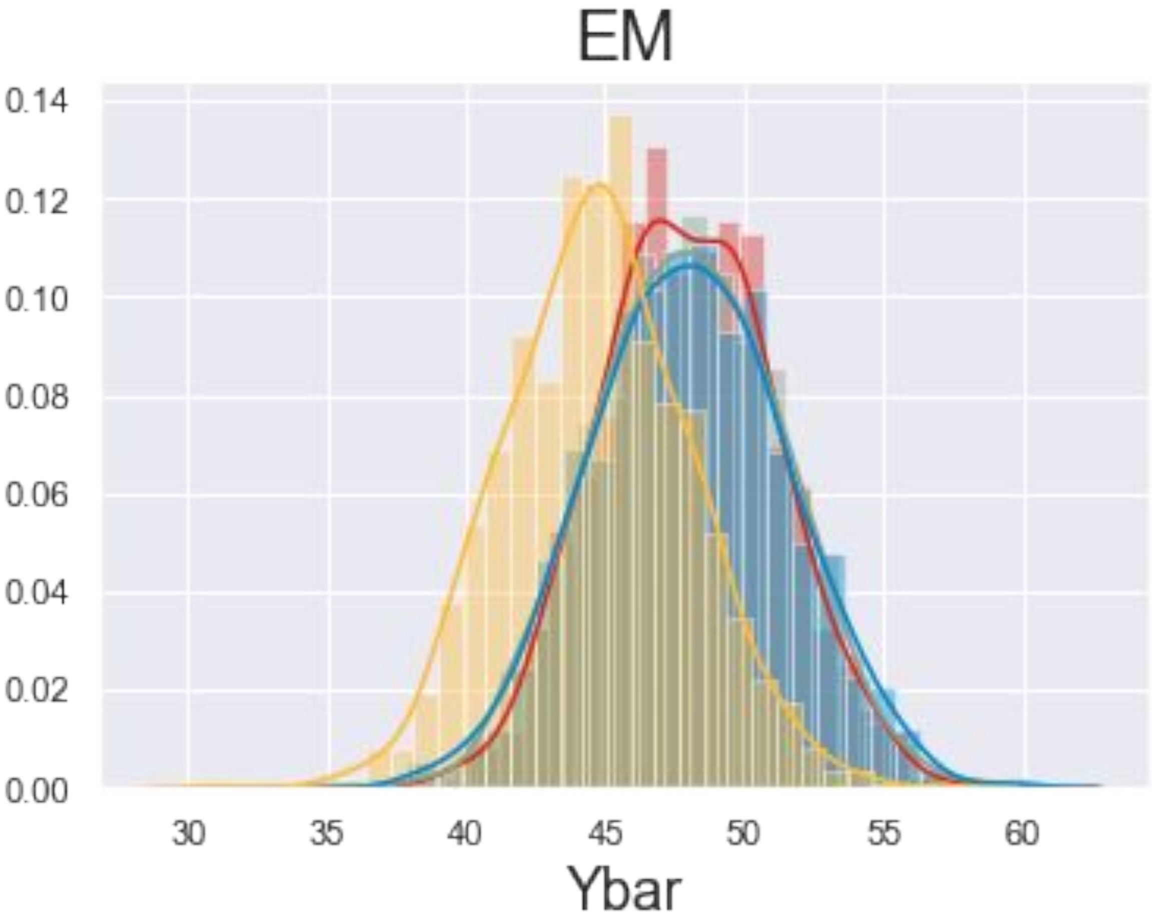
EM Algorithm
Background

1 Initialize $\hat{\mu}$ and $\hat{\Sigma}$

2 E-Step: Estimate missing values

3 M-Step: Maximize (update) $\hat{\mu}$ and $\hat{\Sigma}$

Repeat until $\hat{\mu}$ and $\hat{\Sigma}$ converges



Dataset	Mean	Variance
Full	47.97	10.06
MCAR	47.98	11.52
MAR	48.02	12.41
MNAR	44.70	11.00

04
Conclusion

Techniques that Yield Unbiased Estimates			
	MCAR	MAR	MNAR
Listwise Deletion	X		
Mean Imputation	X		
Regression Imputation	X		
Multiple Imputation	X	X	
Maximum Likelihood Estimation (EM)	X	X	

Conclusions

- The missingness mechanism must be considered before applying techniques

- Testing for missingness mechanism

Conclusions

- Deletion techniques
  - smaller dataset → loss of statistical power
- Single imputation techniques
  - certainty of imputed values = certainty of observed values

Conclusions

- Multiple Imputation (MI) vs. Maximum Likelihood (ML)
  - MI requires many decision points:
    - imputation technique
    - variables included in imputation
    - magnitude of variance for residual term
    - number of datasets to impute
  - Uncertainty in MI regression coefficients
  - ML yield same results each time

Final Thoughts

- Packages available in R / Python
- Consider how to minimize missing data in collection phase
- If MCAR/MAR cannot be assumed:
  - impute using a conservative value
  - adjust inference statements

Thank you!

References

Allison, Paul. *Handling Missing Data by Maximum Likelihood*. SAS Global Forum, 2012.

Don, Yiran and Peng, Chao-Ying Joanne. *Principled Missing Data Methods for Researchers*. Springerplus, 2013.

Enders, Craig. *Applied Missing Data Analysis*. The Guilford Press, 2010

Little, Roderick and Rubin, Donald. *Statistical Analysis with Missing Data*. Wiley, 2019