Handling Missing Data

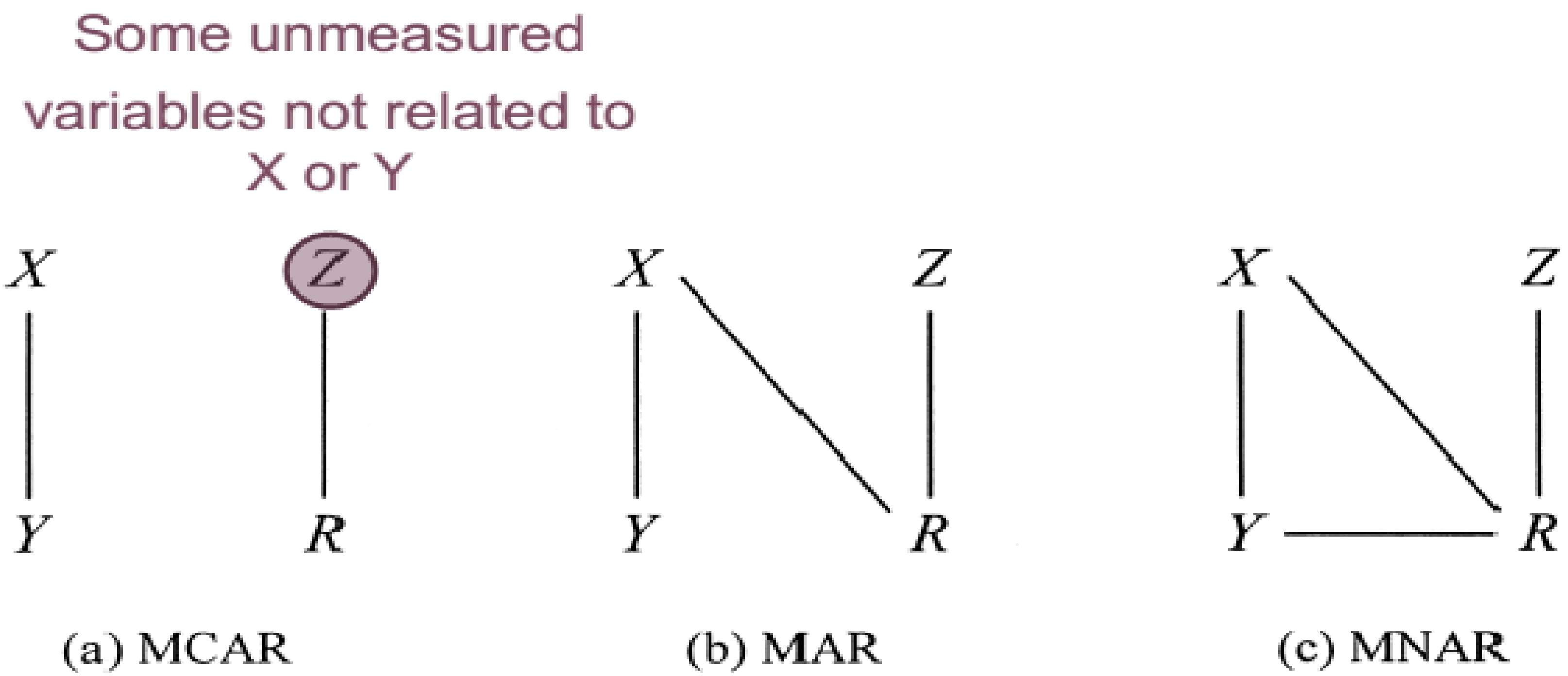
Types of missing data

- Missing Completely At Random (MCAR)
- Missing At Random (MAR)
- Missing Not At Random (MNAR)

OK

PROBLEM

Types of missing data



MCAR: Missingness does not depend on data

MAR: Missingness depends only on observed data

MNAR: Missingness depends on missing data

Types of missing data:Example

Blood Pressure data of 30 participants in January (X) and February (Y)

	Y			
X	Complete	MCAR	MAR	MNAI
Data for individual participants				
169	148	148	148	148
126	123	—	—	—
132	149	—	—	149
160	169	—	169	169
105	138	—	—	—
116	102	—	—	—
125	88	—	—	—
112	100	—	—	—
133	150	—	—	150
94	113	—	—	—
109	96	—	—	—
109	78	—	—	—
106	148	—	—	148
176	137	—	137	—
128	155	—	—	155
131	131	—	—	—
130	101	101	—	—
145	155	—	155	155
136	140	—	—	—
146	134	—	134	—
111	129	—	—	—
97	85	85	—	—
134	124	124	—	—
153	112	—	112	—
118	118	—	—	—
137	122	122	—	—
101	119	—	—	—
103	106	106	—	—
78	74	74	—	—
151	113	—	113	—

- MCAR: Delete 23 Y values randomly
- MAR: Keep Y only where X > 140 (follow-up)
- MNAR: Record Y only where Y > 140 (test everybody again but only keep values of critical participants)

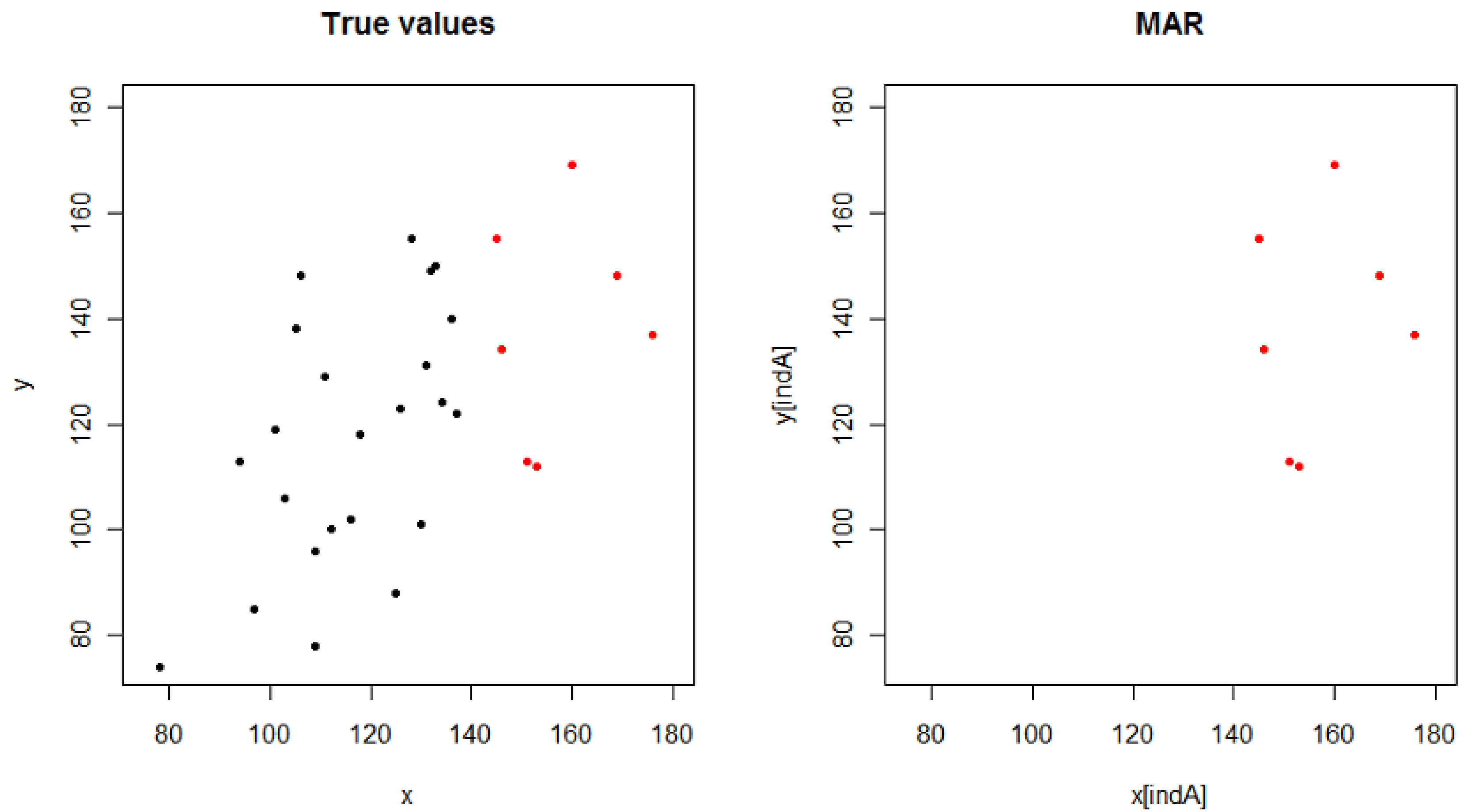
How do you handle missing data?

- Type is not testable.
- Pragmatic:
  - Don’t use methods which hold only in MCAR
  - Use methods which hold in MAR
    - Complete-case analysis(valid for MCAR)
    - Single Imputation(valid for MAR)
    - Multiple Imputation(valid for MAR)

Example Continued

X	Y Complete	MCAR	MAR	MNAR
Data for individual participants				
169	148	148	148	148
126	123	—	—	—
132	149	—	—	149
160	169	—	169	169
105	138	—	—	—
116	102	—	—	—
125	88	—	—	—
112	100	—	—	—
133	150	—	—	150
94	113	—	—	—
109	96	—	—	—
109	78	—	—	—
106	148	—	—	148
176	137	—	137	—
128	155	—	—	155
131	131	—	—	—
130	101	101	—	—
145	155	—	155	155
136	140	—	—	—
146	134	—	134	—
111	129	—	—	—
97	85	85	—	—
134	124	124	—	—
153	112	—	112	—
118	118	—	—	—
137	122	122	—	—
101	119	—	—	—
103	106	106	—	—
78	74	74	—	—
151	113	—	113	—

Example Continued



Black points are missing (MAR)

Complete-case analysis

- Delete all rows, that have a missing value
- Problem:
  - waste of information; inefficient
  - introduces bias if MAR
- OK, if 95% or more complete cases

Single Imputation

- Unconditional Mean Imputation
- Hotdeck Imputation
- Model based Imputation

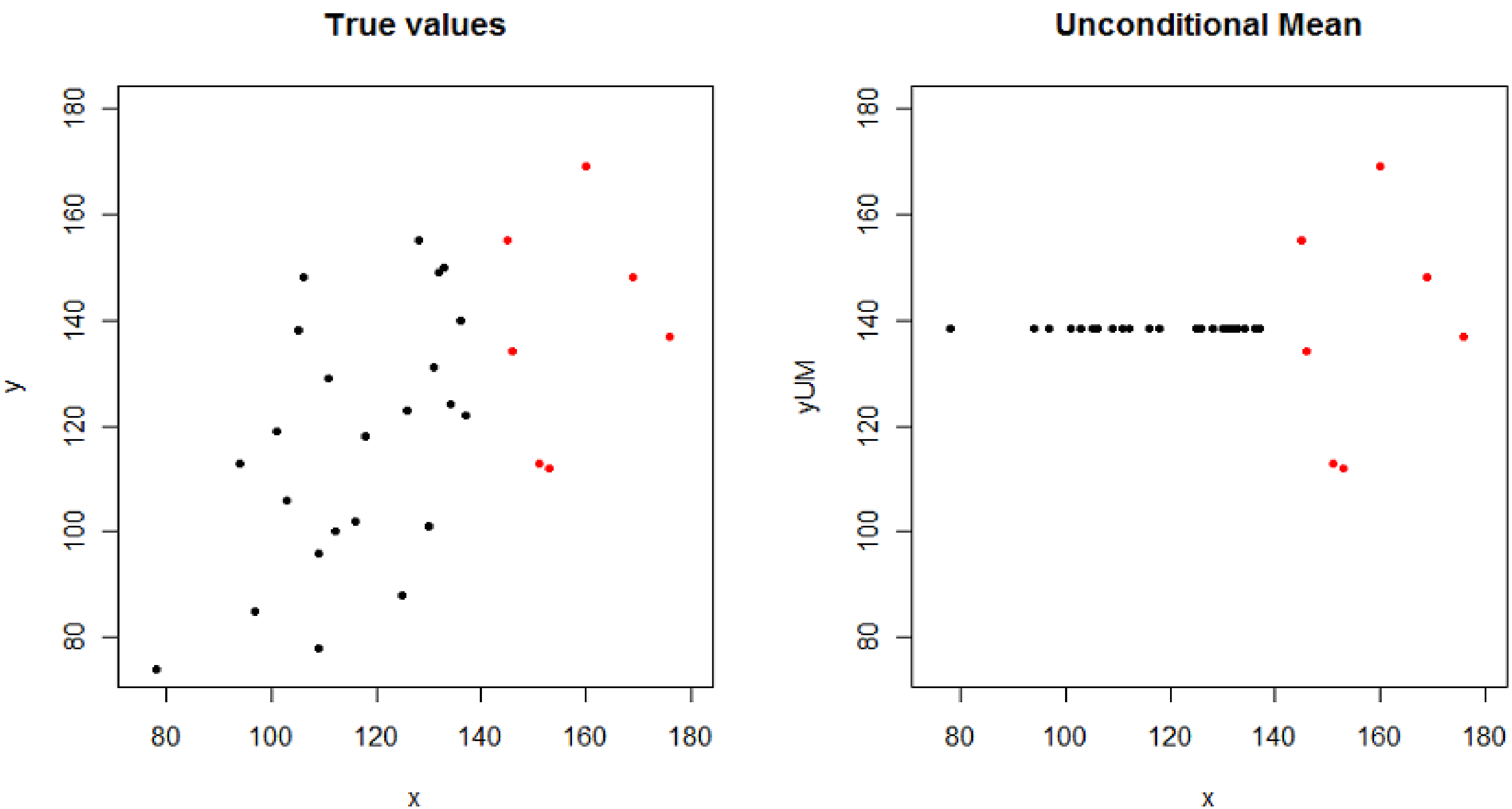
Unconditional Mean Imputation

A	B	C
2.1	6.2	3.2
3.4	3.7	6.3
4.1	4.5	NA

Mean = 4.75

A	B	C
2.1	6.2	3.2
3.4	3.7	6.3
4.1	4.5	4.75

Whats wrong with Unconditional Mean?



+ Mean of Y ok

- Variance of Y wrong

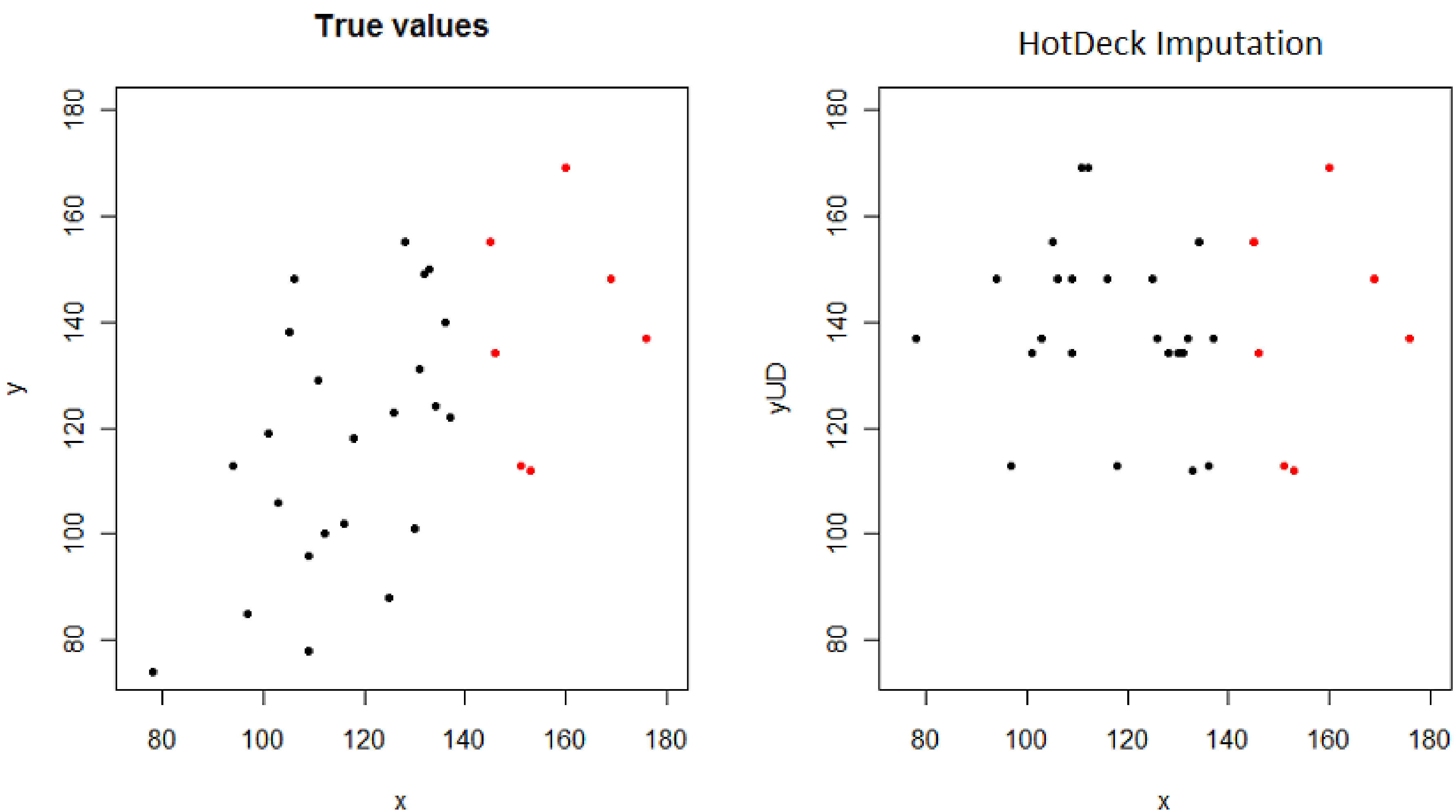
Hotdeck Imputation

A	B	C
2.1	6.2	3.2
3.4	3.7	6.3
4.1	4.5	NA

Randomly select observed value in column

A	B	C
2.1	6.2	3.2
3.4	3.7	6.3
4.1	4.5	6.3

Whats wrong with HotDeck Imputation?



+ Mean of Y ok, Variance better

- Correlation btw X and Y wrong

Model based Imputation(Linear Regr)

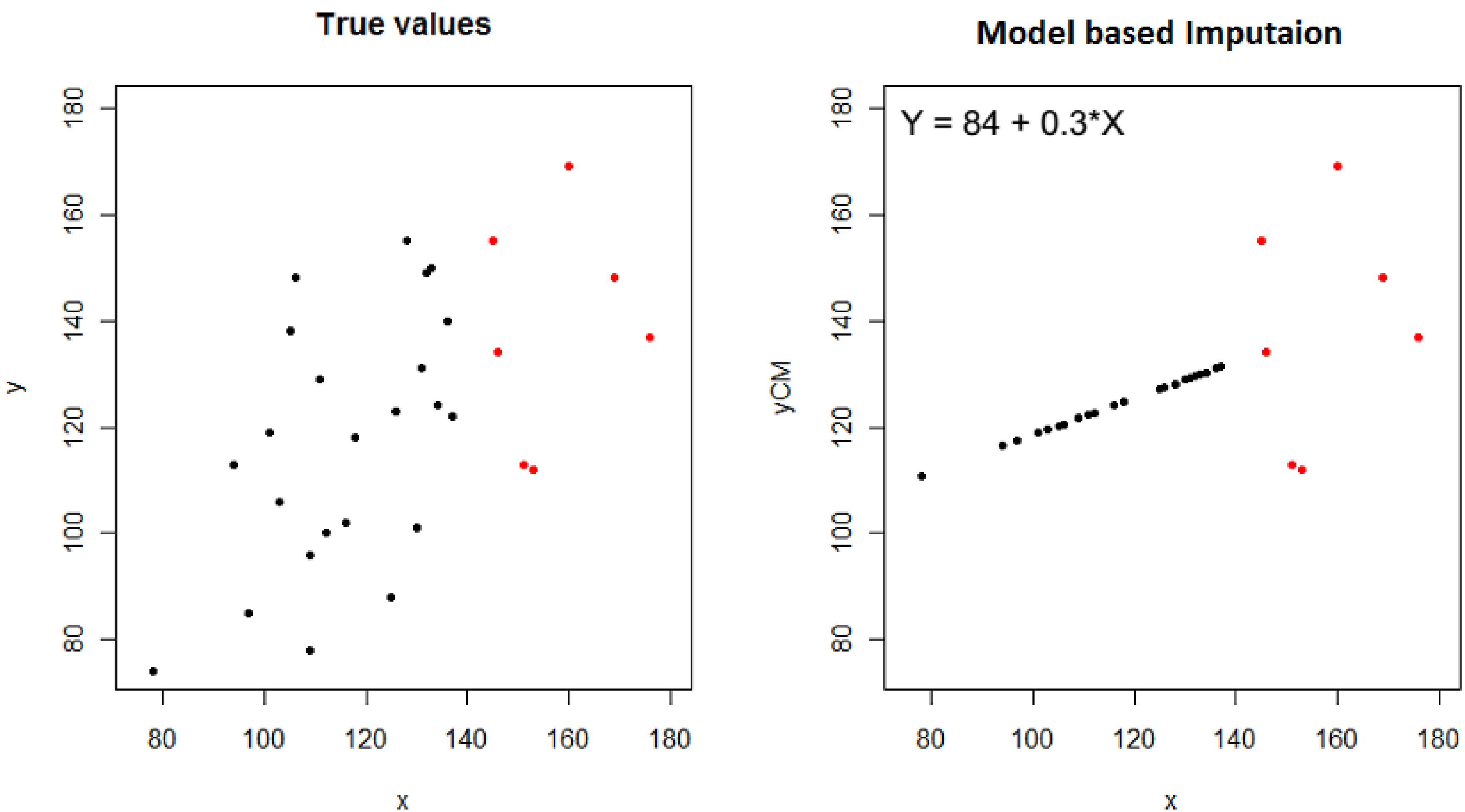
- Build model with C as target variable and A & B as features
- Fill the missing data of C with the predictions of learned model

A	B	C
2.1	6.2	3.2
3.4	3.7	6.3
4.1	4.5	NA

Prediction of **model**

A	B	C
2.1	6.2	3.2
3.4	3.7	6.3
4.1	4.5	8

Whats wrong with Model based Imputation?



+ Conditional Mean of Y ok

+ Correlation ok

- (Conditional) Variance wrong

Model based Imputation(Linear Regr + Noise)

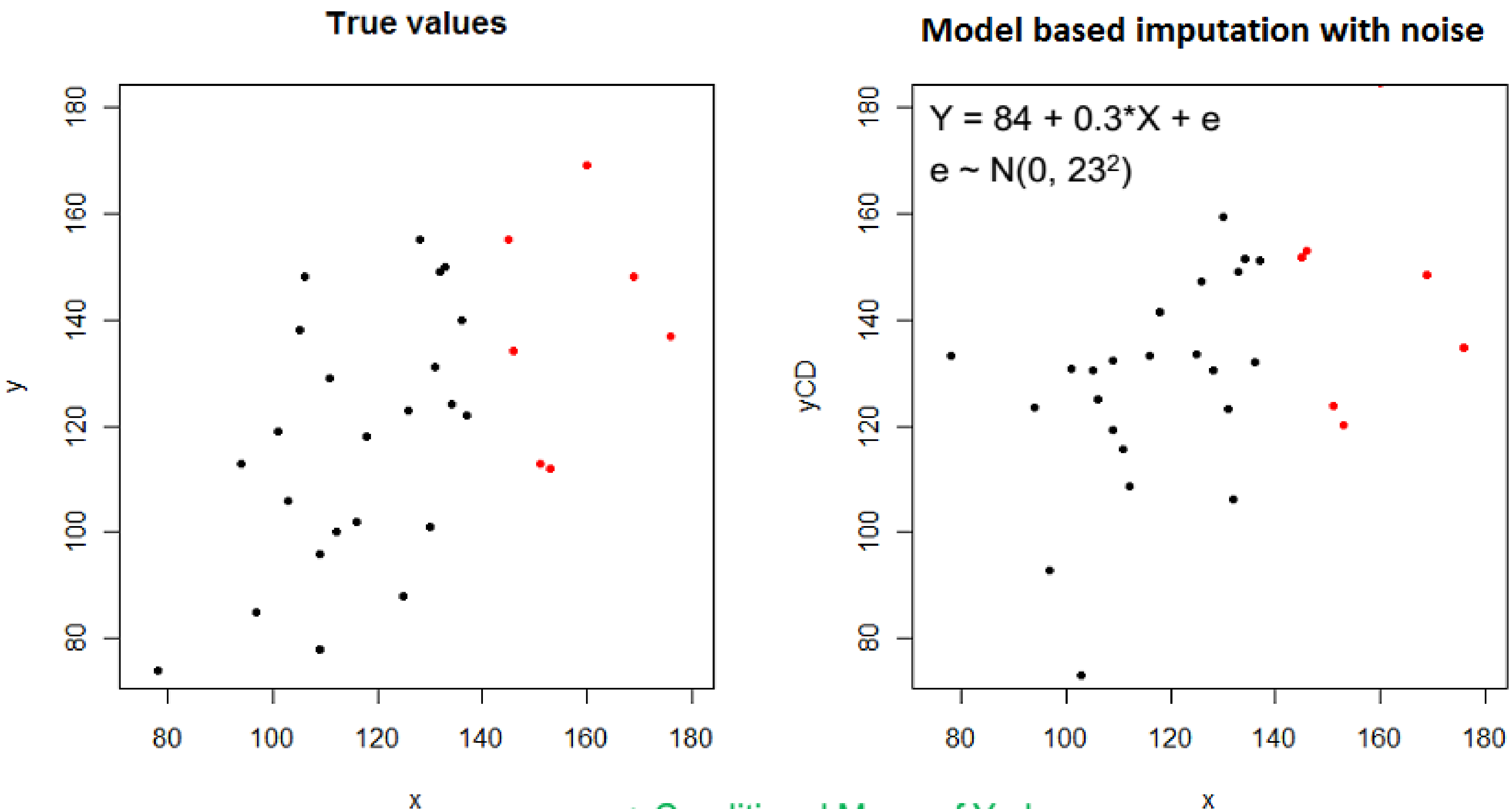
- Build model with C as target variable and A & B as features
- Fill the missing data of C with the predictions of learned model plus noise

A	B	C
2.1	6.2	3.2
3.4	3.7	6.3
4.1	4.5	NA

Prediction of
model
PLUS NOISE

A	B	C
2.1	6.2	3.2
3.4	3.7	6.3
4.1	4.5	8.3

Whats wrong with Model based Imputation?

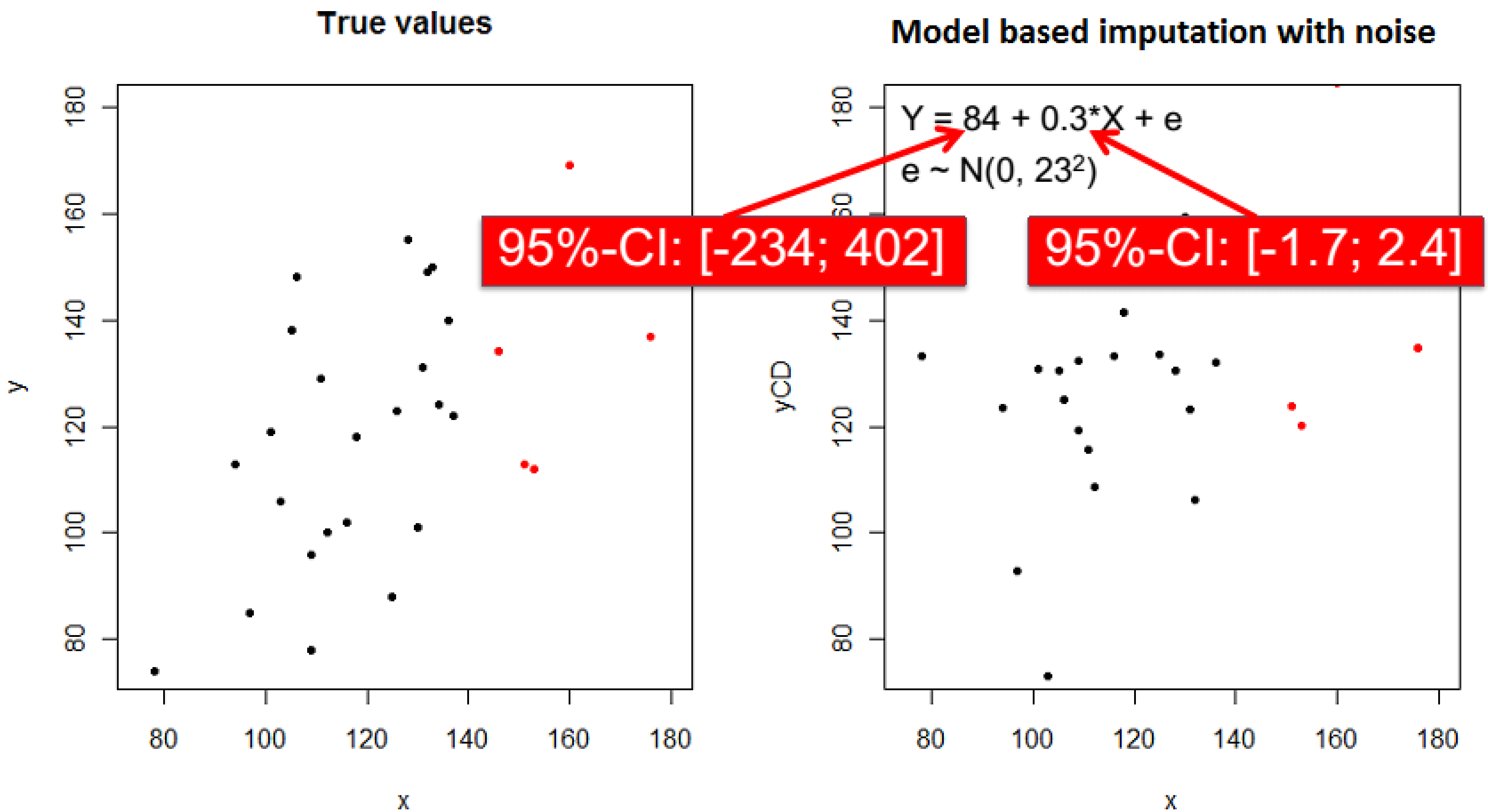


+ Conditional Mean of Y ok

+ Correlation ok

+ Conditional Variance of Y ok

Whats wrong with Model based Imputation?



Problem: We ignore uncertainty

Model based Imputation(RF)

- Good trade-off between ease of use/accuracy
- Works with mixed data types(categorical & continuous)
- Estimates the quality of imputation with OOB error
    - OOB error: Imputation error as percentage of total variation

      close to 0: good

      close to 1 : bad

Model based Imputation(RF)

A	B	SEX
2.1	NA	M
3.4	3.7	F
4.1	4.5	NA

Fill in Random values

A	B	SEX
2.1	3.0	M
3.4	3.7	F
4.1	4.5	F

Model based Imputation(RF)

Step1:

A	B	SEX
2.1	3.0	M
3.4	3.7	F
4.1	4.5	F

Learn B ~ A + SEX
with Random Forest

Apply B ~ A + SEX → update value

Model based Imputation(RF)

Step2:

A	B	SEX
2.1	3.2	M
3.4	3.7	F
4.1	4.5	F

Learn SEX ~ A + B
with Random Forest

Apply SEX ~ A + B → update

Model based Imputation(RF)

- Repeat steps 1 & 2 until some stopping criterion is reached (no real convergence; stop if updates start getting bigger again)

Pros & Cons of rfImpute

- Effects are OK, if MAR holds
- Estimation of imputation error
- Accuracy might be too optimistic, because
    - imputed values have no random scatter
    - model for prediction was taken to be the true model, but it is just an estimate

Measuring quality of Imputation

- Normalized Root Mean Squared Error (NRMSE):

$$NRMSE = \sqrt{\frac{mean(Y_{com} - Y_{imputed})^2}{var(Y_{com})}}$$

- Proportion of falsely classified entries (PFC) over all categorical values

$$PFC = \frac{nmb.\ missclassified}{nmb.\ categorical\ values}$$

Multiple Imputation

- The imputed values by Single Imputation is too optimistic. It ignores uncertainty in model parameters

- Multiple Imputation incorporates both
    - residual error
    - model uncertainty

Multiple Imputation: Intuitive Idea

- Fill in random values
- Iteratively predict values for each variable until some convergence is reached (as in randomForest)
- Sample values for residuals AND for model parameters. Usually, Gibbs sampler is used.