Data Wrangling and Data Analysis
Missing Data and Imputation

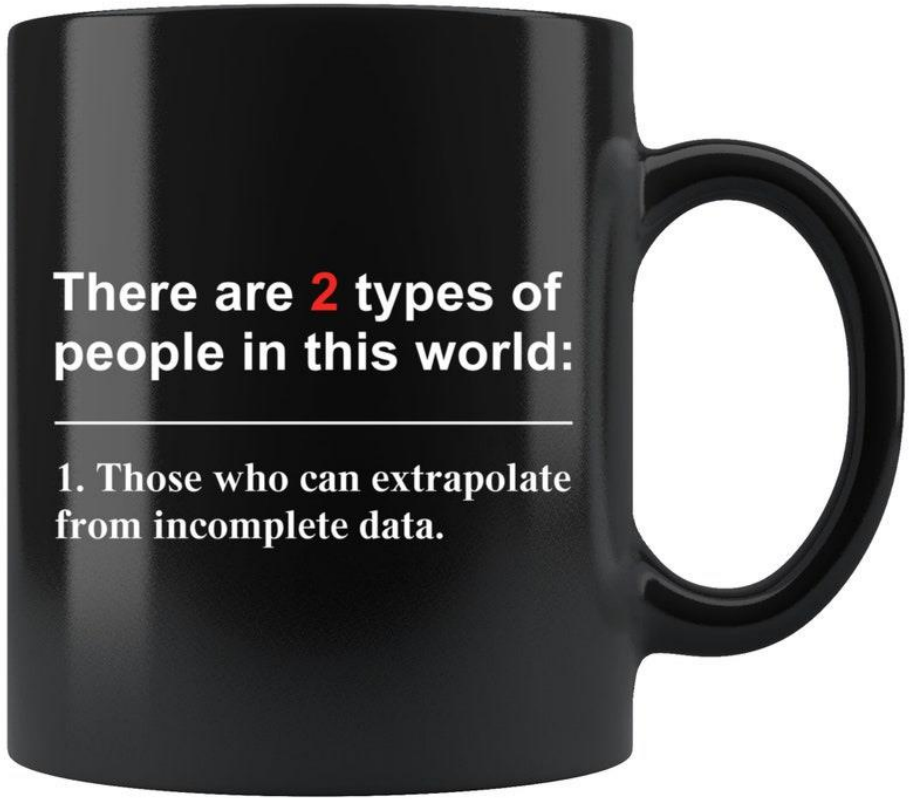
Daniel Oberski

Department of Methodology & Statistics

Utrecht University

This week

- Lecture 1: Problems with missing data
- **Lecture 2: Solutions to problems with missing data**
- Lecture 3: Clustering



Strategies to deal with missing data

Prevention

(the golden road?)

Imputation

- Ad-hoc methods
- Single imputation
- Multiple imputation

Adjustment

- Weighting methods
- Likelihood methods
- EM-algorithm
- …

Imputation

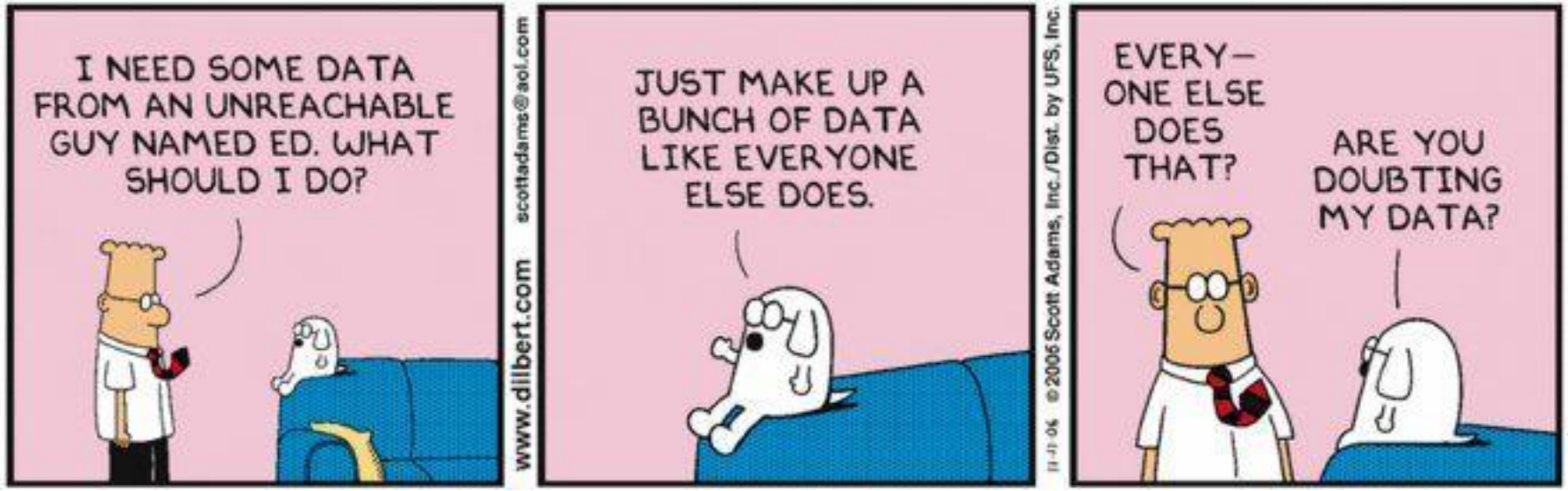
Replacing missing values with *guessed* values

```
##   age weight
## 1 13   42
## 4 14   NA
## 6 18   61
## 5 23   70
## 3 24   73
## 7 25   68
## 2 40   80
```

```
##   age weight
## 1 13   42
## 4 14   47
## 6 18   61
## 5 23   70
## 3 24   73
## 7 25   68
## 2 40   80
```

Imputation

- Yields “complete” data
- Statistics are now defined
- Hooray?



Good imputations are *plausible*

https://huggingface.co/black-forest-labs/FLUX.1-Fill-dev

Deductive imputation

- If we know height and weight, we can calculate BMI
- A male child is probably not pregnant

Converse also holds: If an *observed* value “must” be wrong, we can correct it or *make* it missing

- Example: database may contain a 14 yr old female with 3 kids, married for 20 years and working as a manager at a public school
- This may be a data entry error (perhaps it should have been 41?). We can set this value missing and let our algorithms do the rest.

Missing data procedures and your (implicit) assumptions

- Missing data important **because they can bias the analysis**
- Bias depends on NDD/SDD/UDD

→ **Goal: prevent the bias**

- **Logic of all missing data procedures:**
  - IF NDD/SDD is causing the bias,
  - THEN procedure *xyz* will remove that bias
- (On very rare occasions, specific forms of UDD can be dealt with)

- In other words, each procedure has an **implicit assumption**

Methods and their assumptions

	Assumption needed to get unbiased estimate			
	Mean	Regression coef.	Correlation	Standard Error
Listwise deletion				
Mean imputation				
Regression imputation				
Stochastic imputation				

Listwise deletion

Also known as:

- “Complete Case Analysis”
- “we deleted missing data”
- “huh? Whaddayamean missing data?”

Listwise deletion

```
##   age weight
## 1 13   42
## 4 14   NA
## 6 18   61
## 5 23   70
## 3 24   73
## 7 25   68
## 2 40   80
```

```
##   age weight
## 1 13   42
## 6 18   61
## 5 23   70
## 3 24   73
## 7 25   68
## 2 40   80
```

Listwise deletion

Advantages

- Simple (default in most software)
- Unbiased under NDD

Disadvantages

- Large loss of information, hopeless with many features
- Biased under SDD and UDD, even for simple stuff like mean

Mean imputation

Replace the missing values by the mean of the observed data

Advantages

- Simple
- Unbiased for the mean, under NDD

Disadvantages

- Doesn’t work
- (okay, except in some very specific circumstances)

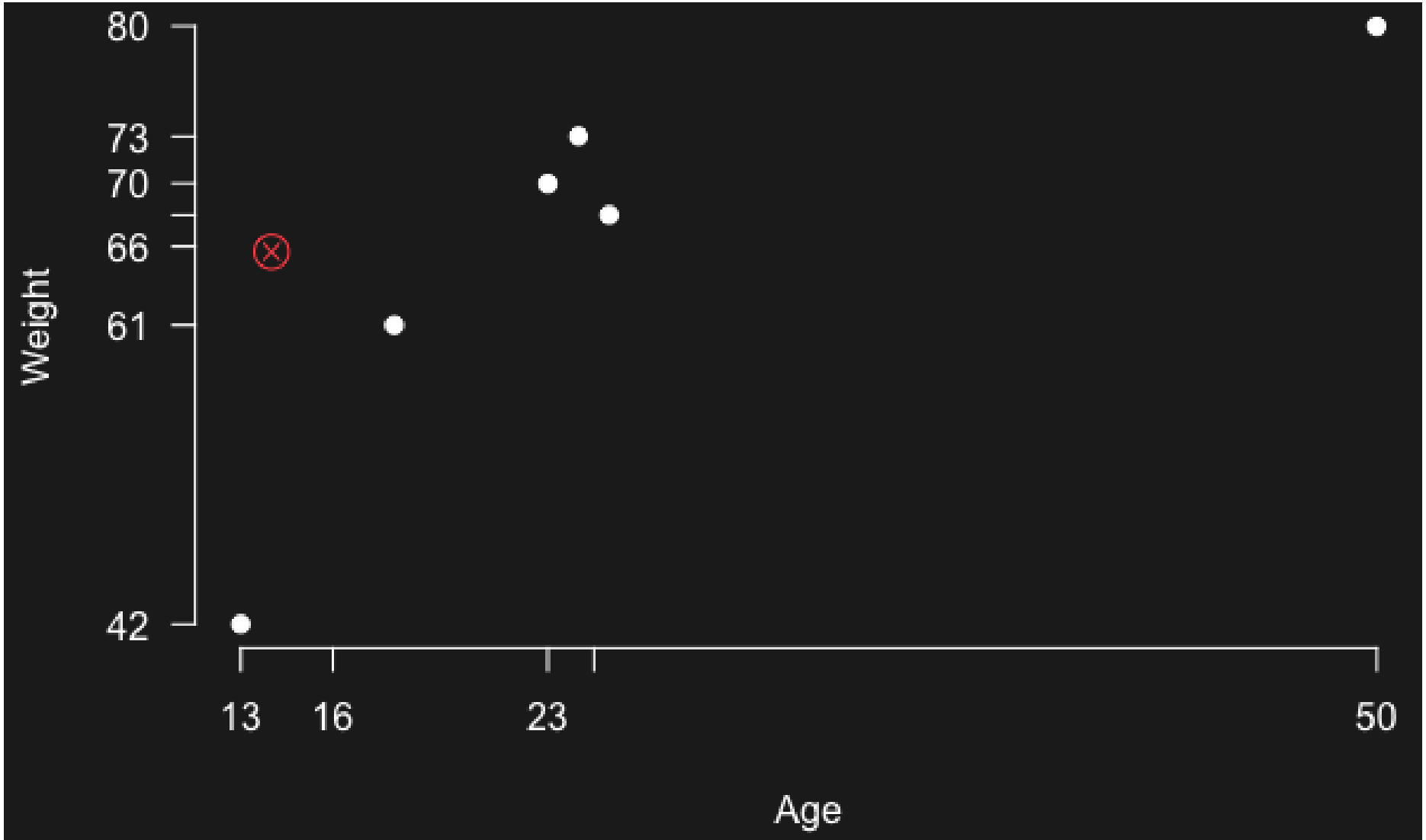
Mean imputation

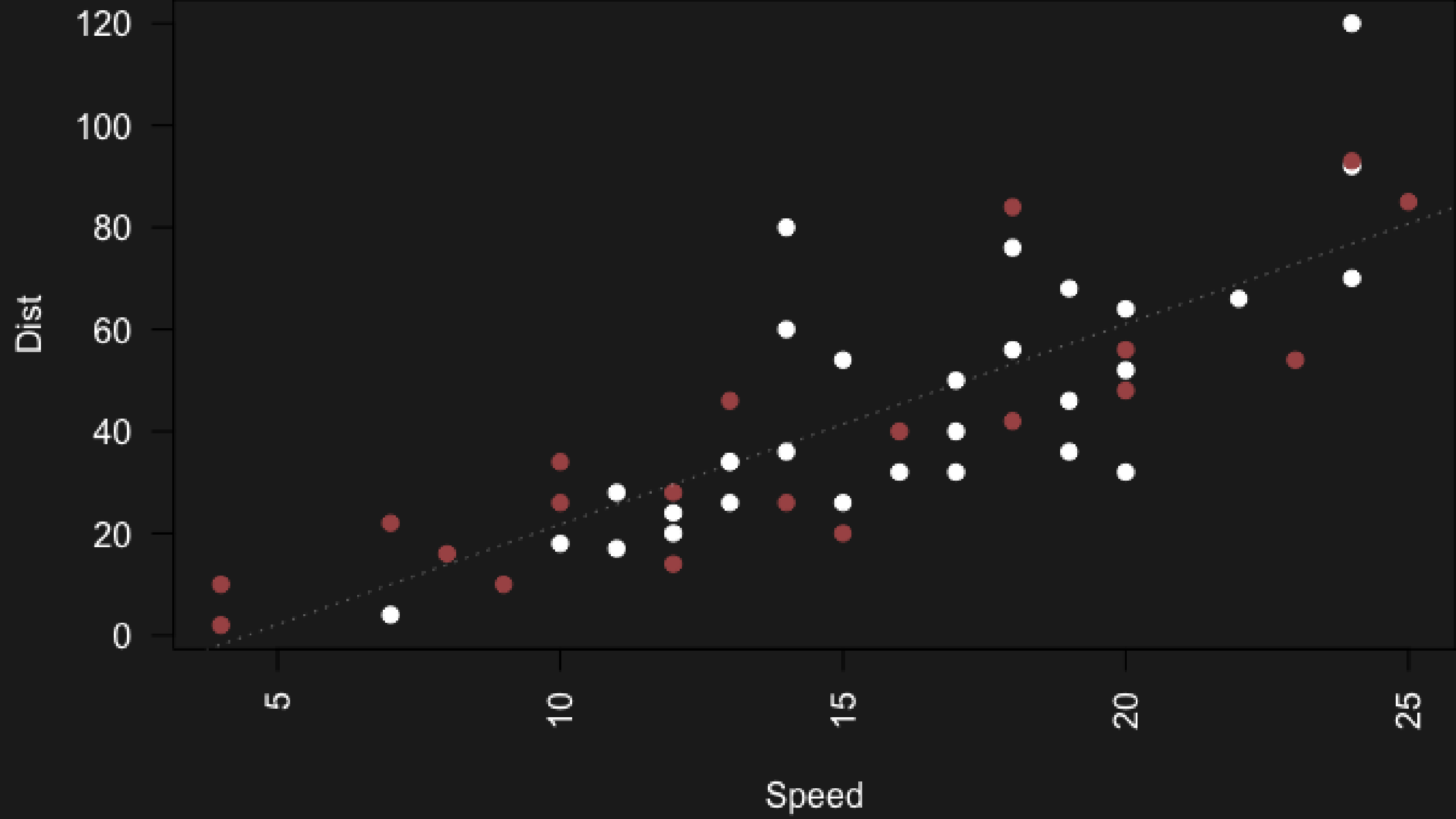
```
##   age weight
## 1 13   42
## 4 14   NA
## 6 18   61
## 5 23   70
## 3 24   73
## 7 25   68
## 2 40   80
```

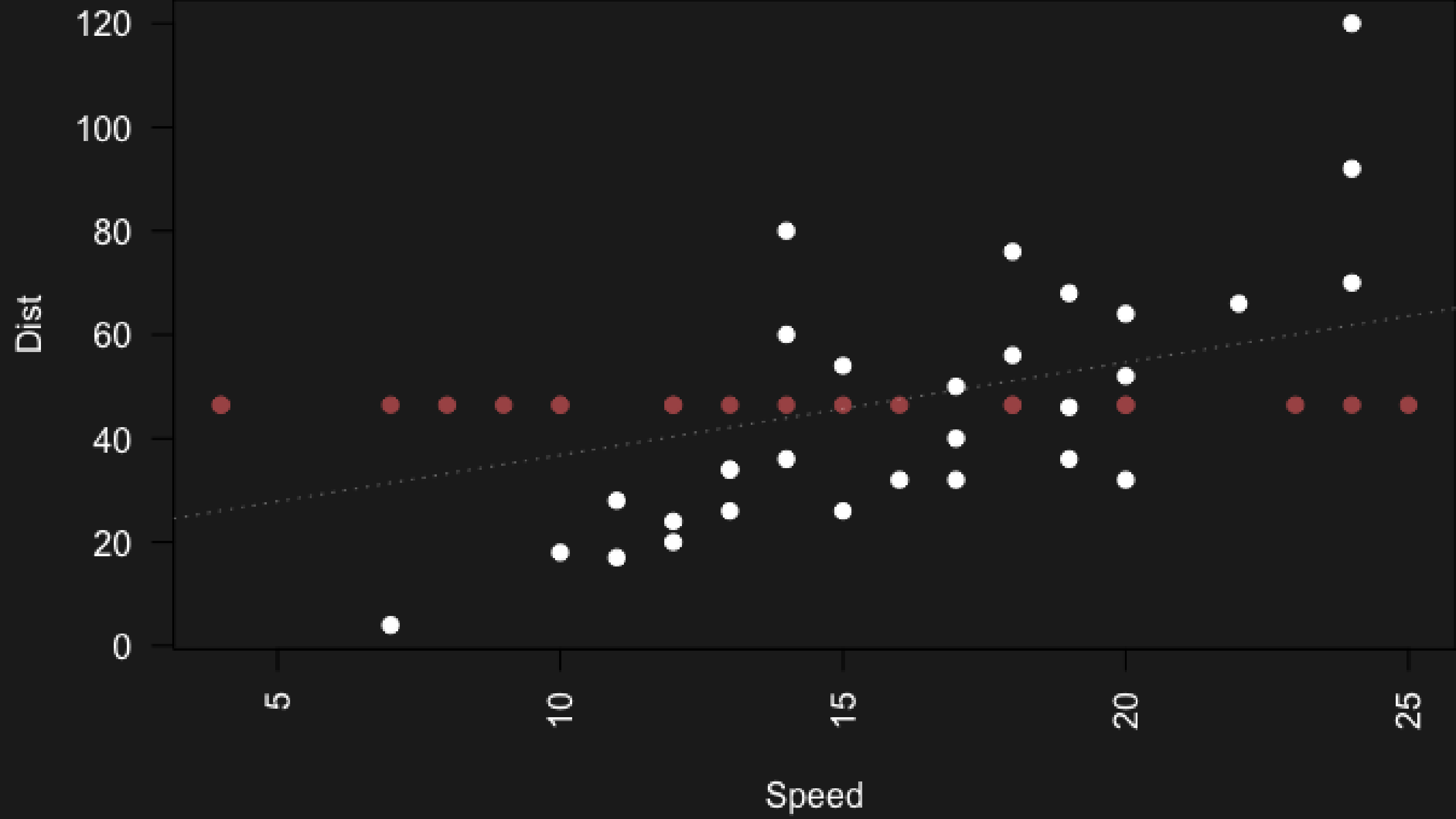
```
##   age weight
## 1 13   42
## 4 14   65.667
## 6 18   61
## 5 23   70
## 3 24   73
## 7 25   68
## 2 40   80
```

Mean imputation: univariate perspective

Mean imputation: bivariate perspective

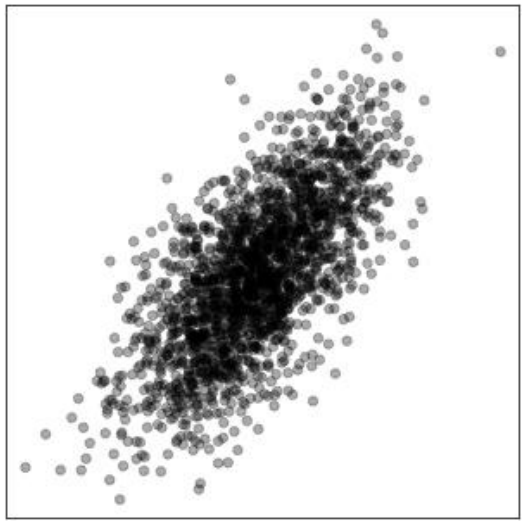



120
100
80
60
40
20
0
Dist
5
10
15
20
25
Speed


120
100
80
60
40
20
0
Dist
5
10
15
20
25
Speed

Mean imputation in prediction models

Before missingness



$\mathrm{var}(y) = 2$

$\mathrm{MSE}(y|x) = 1$

After mean imputation

1. What is the **cross-validated MSE** of the best model after mean imputation?
2. What is the **true MSE** of this model for new data?

Mean imputation

Disadvantages

- Disturbs the distribution
- Underestimates the variance
- Biases correlations to zero
- Biased under SDD

AVOID

(unless you know what you are doing (and probably even then))

Intermezzo: tiny crash course/reminder of classical statistics

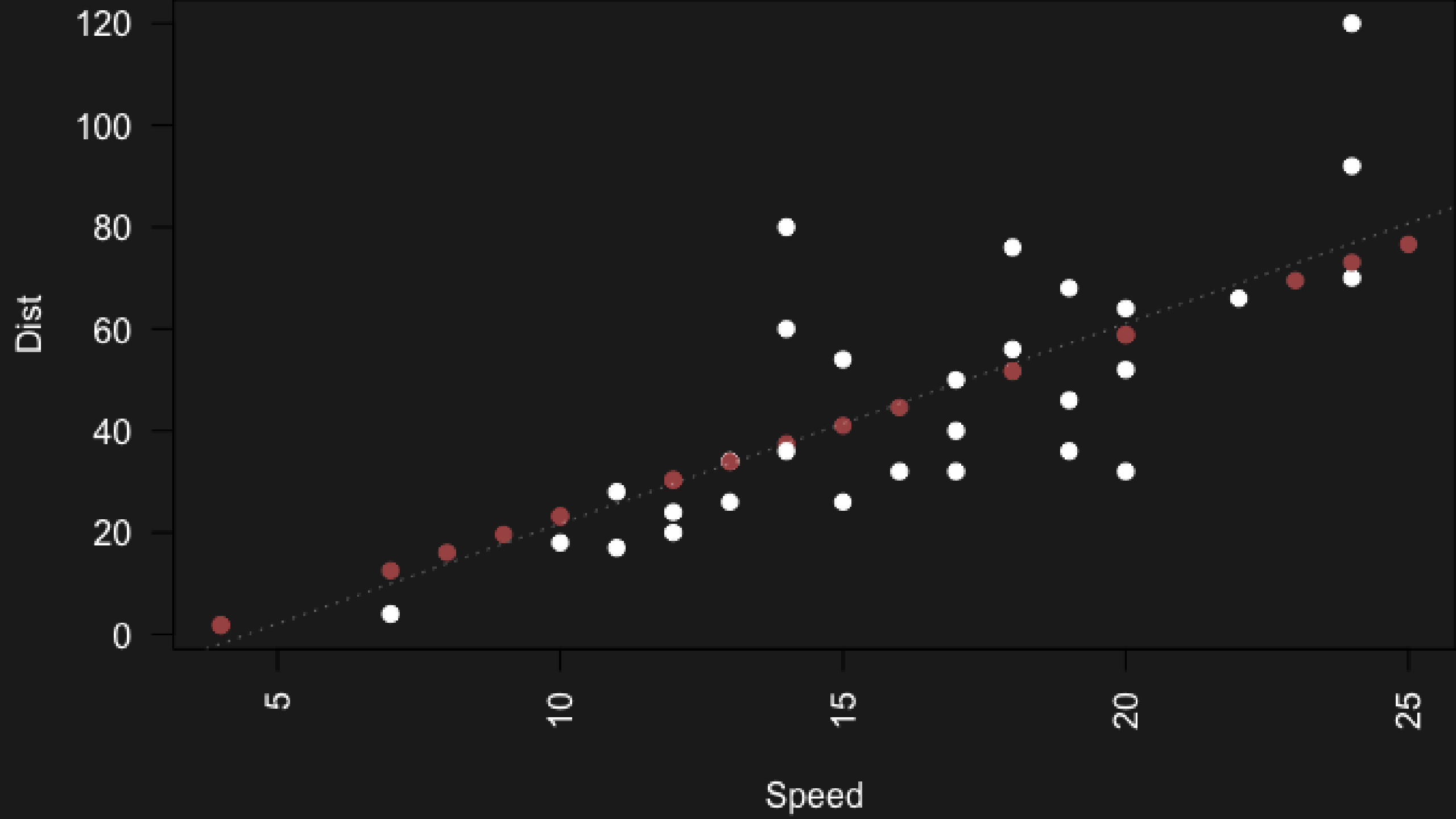
“Regression” imputation (in the statistics sense)

- Just *predict* the missing value
- Fit model for weight under listwise deletion: the **imputation model**
- Predict body weight for records with missing weight
- Replace missing values by prediction

Regression imputation

```
##    age weight
## 1 13   42
## 4 14   NA
## 6 18   61
## 5 23   70
## 3 24   73
## 7 25   68
## 2 40   80
```

```
##    age weight
## 1 13   42
## 4 14   53.45
## 6 18   61
## 5 23   70
## 3 24   73
## 7 25   68
## 2 40   80
```

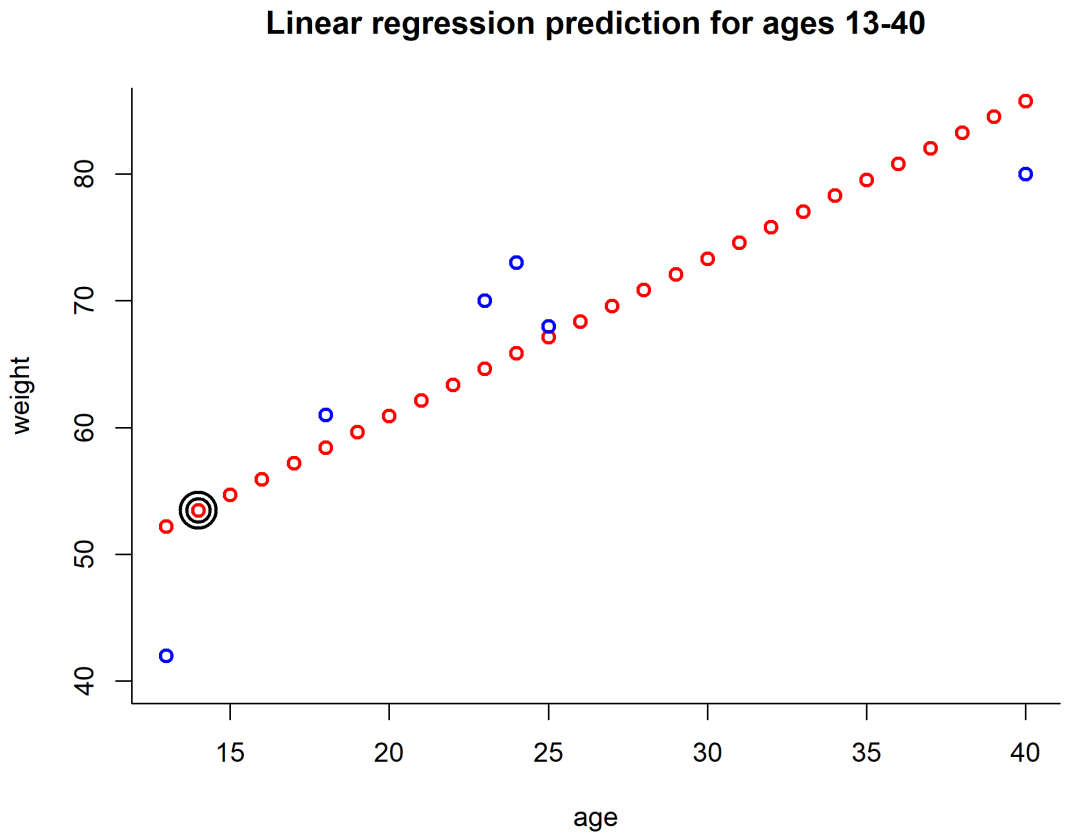

120
100
80
60
40
20
0
Dist
5
10
15
20
25
Speed

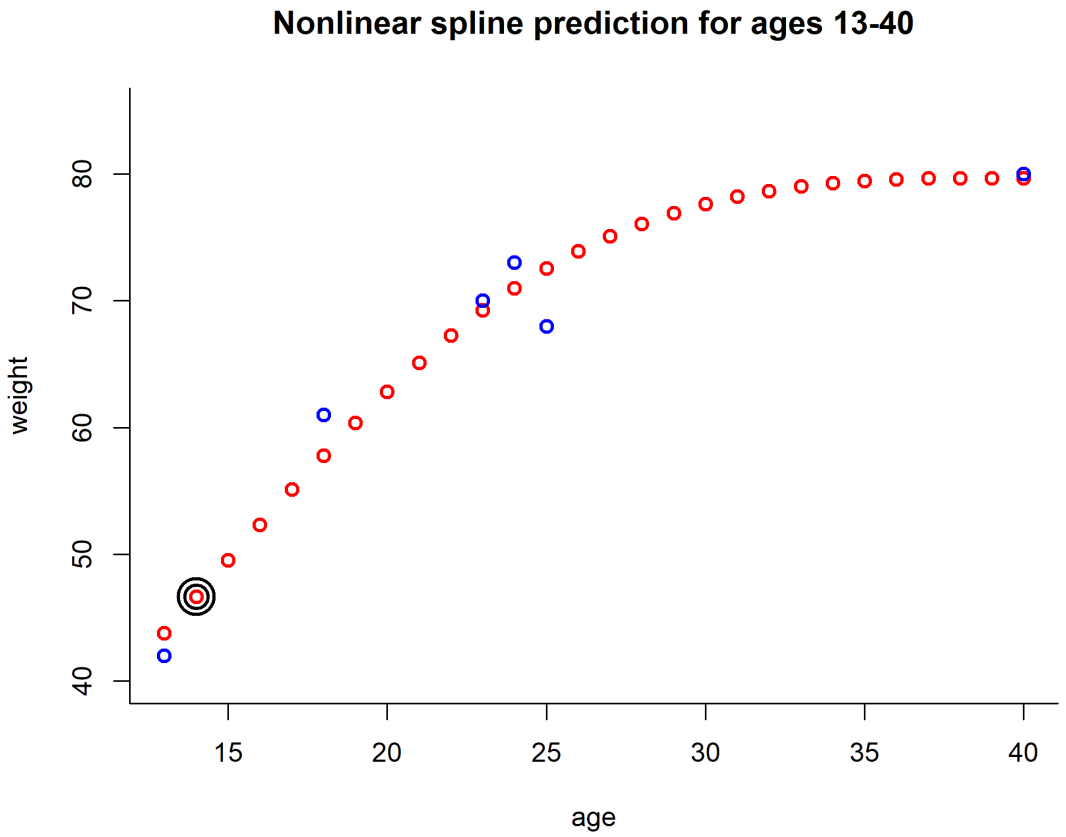
Regression imputation

Advantages

- Unbiased estimates of regression coefficients (under SDD)
- Good approximation to the (unknown) true data if explained variance is high
- The better your prediction, the better your approximation

The better your prediction, the better your approximation





Regression imputation: spline

```
##    age weight
## 1 13   42
## 4 14   NA
## 6 18   61
## 5 23   70
## 3 24   73
## 7 25   68
## 2 40   80
```

```
##    age weight
## 1 13   42
## 4 14   46.65
## 6 18   61
## 5 23   70
## 3 24   73
## 7 25   68
## 2 40   80
```

Regression imputation

Disadvantages:

- Artificially increases correlations
- Consequently: underestimates prediction error
- Systematically underestimates the uncertainty
- *p*-values too optimistic, confidence intervals too narrow

Stochastic regression imputation

- Like regression imputation, but adds noise to the predictions to reflect uncertainty
- Uncertainty in the form of:
  - Parameter uncertainty in the prediction model
  - Uncertainty due to unexplained variance in the target feature

- Related to *prediction intervals* (ISLR sections 3.2.2:“Four” and exercises on pages 111-112)

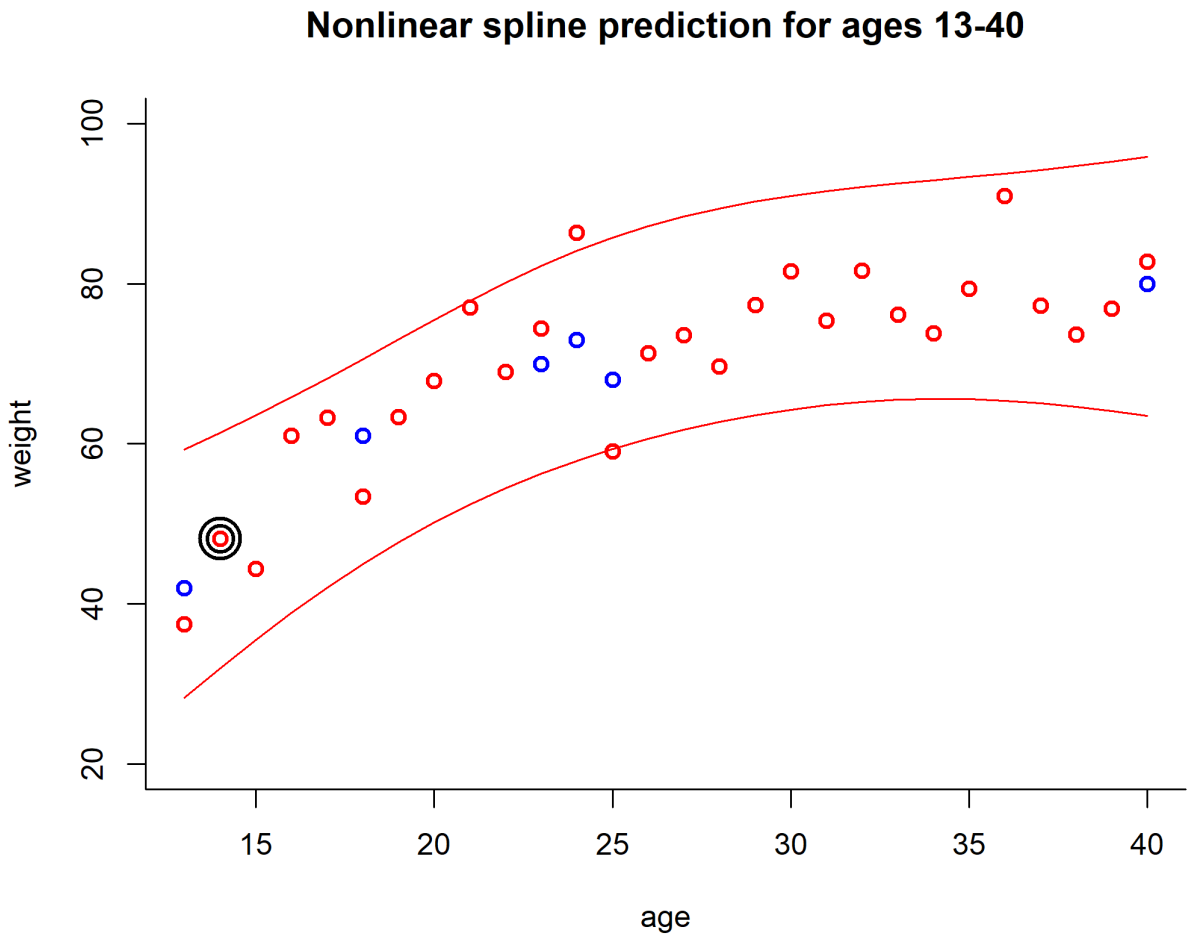
stochastic

/stəˈkastɪk/

adjective **TECHNICAL**

having a random probability distribution or pattern that may be analysed statistically but may not be predicted precisely.

Stochastic regression imputation



Stochastic regression imputation

```
##    age weight
## 1 13   42
## 4 14   NA
## 6 18   61
## 5 23   70
## 3 24   73
## 7 25   68
## 2 40   80
```

```
##    age weight
## 1 13   42
## 4 14   48.13
## 6 18   61
## 5 23   70
## 3 24   73
## 7 25   68
## 2 40   80
```

Stochastic regression imputation

Advantages:

- Preserves the distribution of body weight
- Preserves the correlation between age and weight in the imputed data

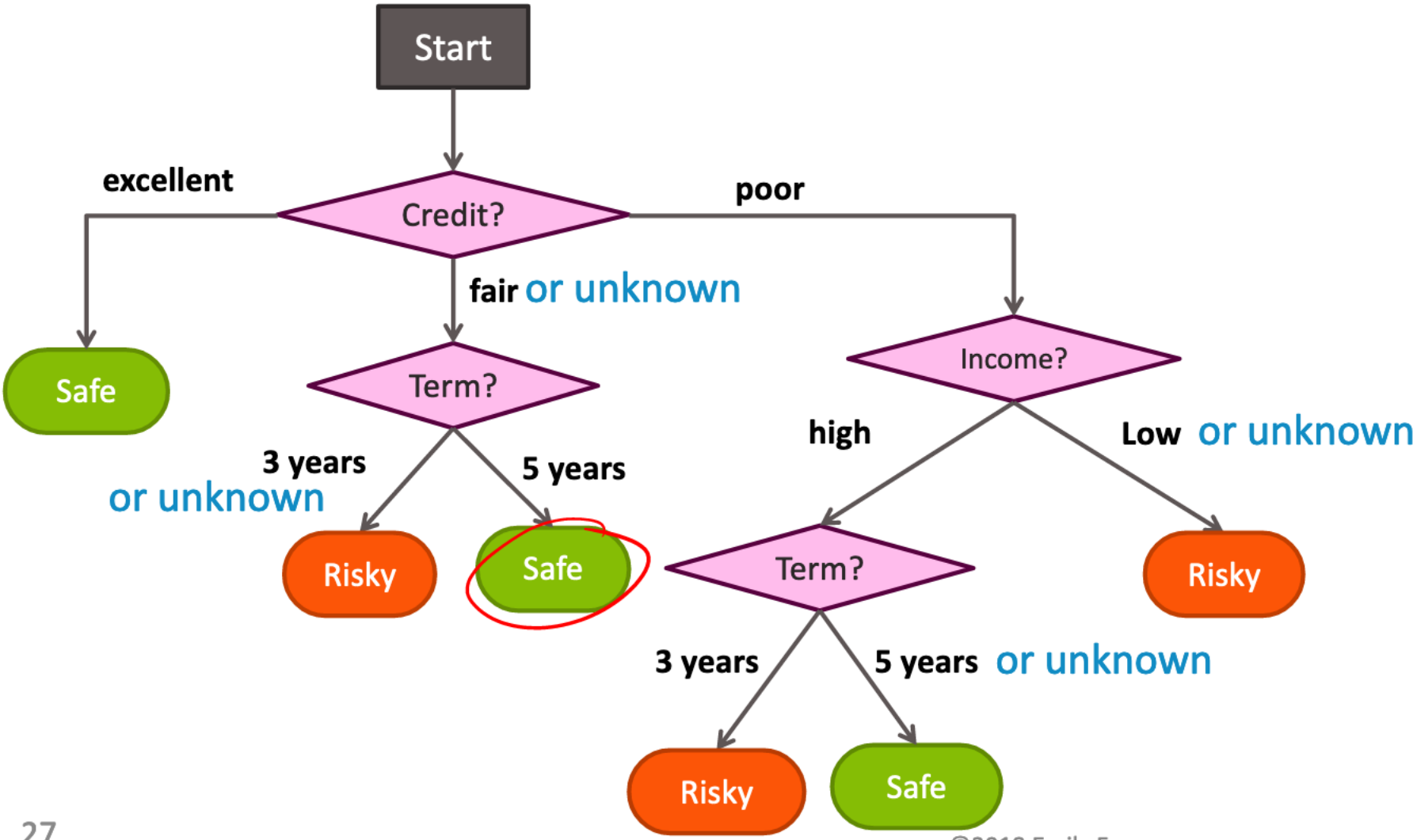
Disadvantages:

- Fiddly
- Single imputation does not take uncertainty imputed data into account, and incorrectly treats them as real

“Embedded” methods (model-based)

- Don’t impute, deal with missing values somehow in the (prediction) model itself
- Depends on the model you are using
- Almost always assumes SDD implicitly
- Example on next slide given with classification tree, but other models may have a different approach

x_i = (Credit = ?, Income = high, Term = 5 years)



- `xgboost` learns "default direction" for missing values
- Tries both possibilities at each node and picks the one with most gain (greedy)

Source: Fox (2018)

Handy overview of methods

	Assumption needed to get unbiased estimates			
	Mean	Regression coef.	Correlation	Standard Error
Listwise deletion	NDD	NDD	NDD	Too large
Mean imputation	NDD	-	-	Too small
Regression imputation	SDD	SDD	-	Too small
Stochastic imputation	SDD	SDD	SDD	Too small

Imputing one value for a missing datum cannot be correct in general, because we don't know what value to impute with certainty (if we did, it wouldn't be missing).

Donald Rubin

Fixing the SE: *Multiple* Imputation

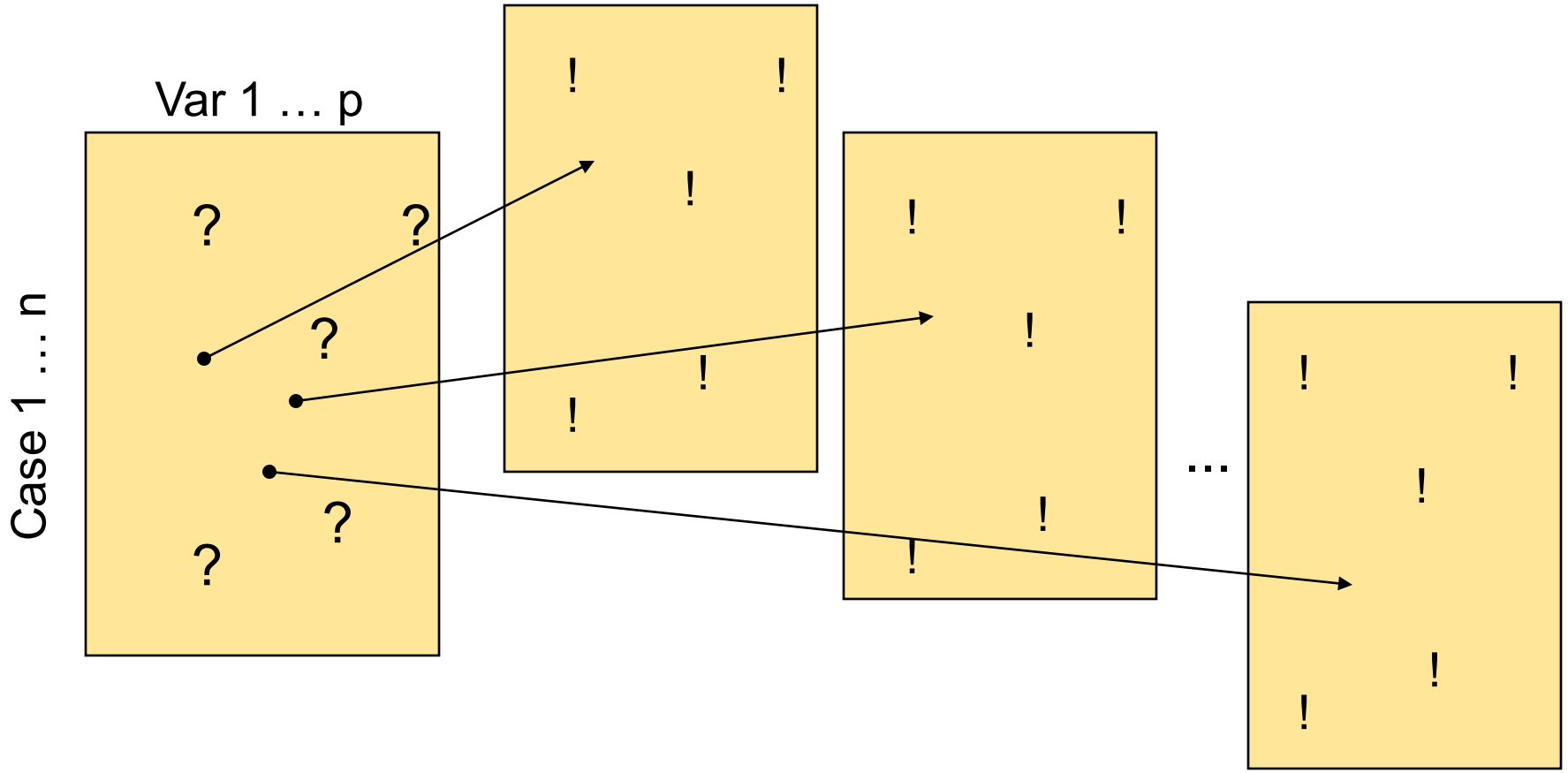
- Stochastic imputation, but create multiple datasets
- Each dataset is slightly different

- Appropriately consider the uncertainty around the imputed value:
  1. Perform analysis multiple times
  2. Pool results

Steps in Multiple Imputation

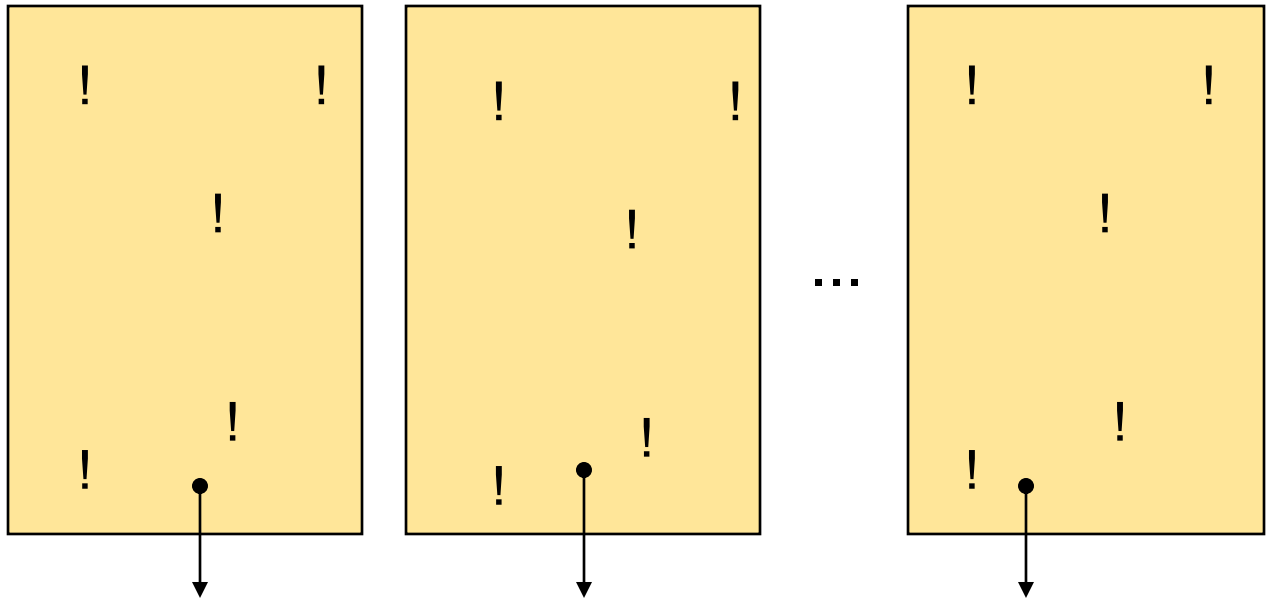
1. Create *m* imputations
2. Analyze *m* completed data sets
3. Combine the *m* results into one

Multiple Imputation: Imputation Step



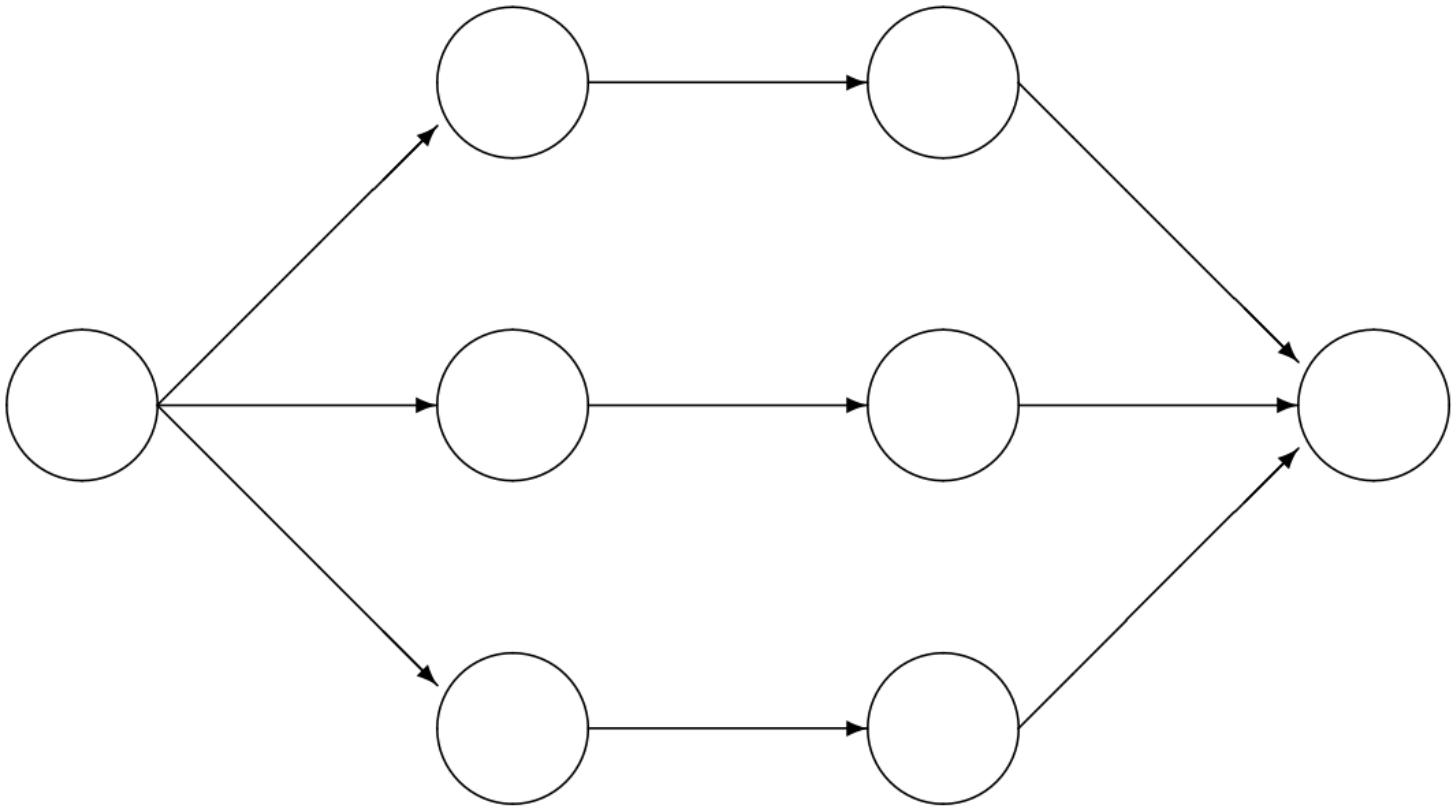
Create *m* different imputed data sets

Multiple Imputation: Analysis Step



Combine results
= take the **mean**,
and calculate **SE** from original SE + dataset variance

Multiple imputation

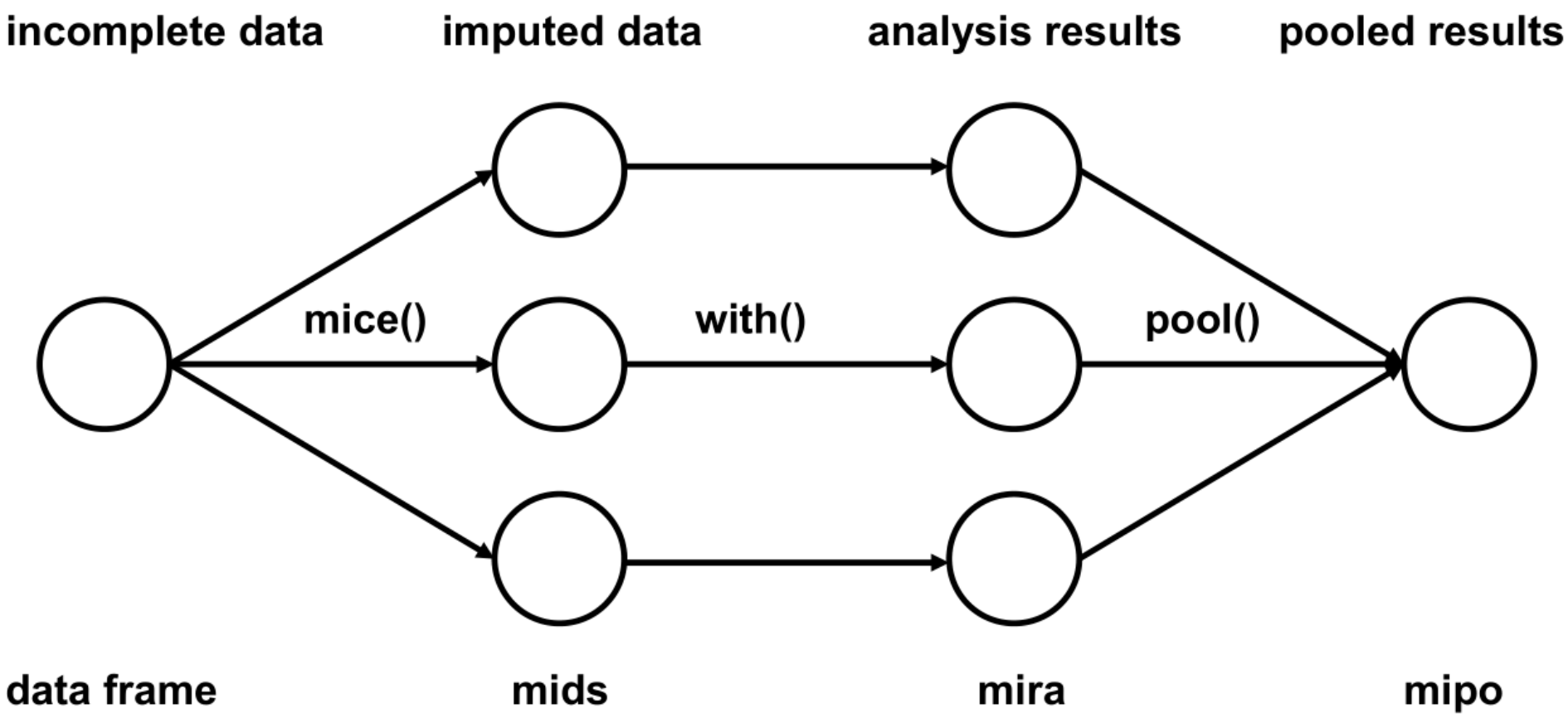


Incomplete data Imputed data Analysis results Pooled results

Software

- The `R` package `mice` performs multiple imputation and automatic pooling of results
- It has support for many analysis methods, such as `anova`, `lm`, `glm`, and many more
- Python `sklearn` has `mice` algorithm in `sklearn.impute.IterativeImputer`
- `sklearn` documentation: “It is still an open problem as to how useful single vs. multiple imputation is in the context of prediction and classification when the user is not interested in measuring uncertainty due to missing values.”

Typical workflow for mice



Ad hoc alternative to pooling: sensitivity

- If conclusion does not change across the *m* imputed datasets, conclusion not sensitive to the imputation
- For your topic this may be enough

Conclusion

- Single imputation does not account for all uncertainty
- One solution is multiple imputation
- Analysis needs to be pooled after being performed on multiply imputed datasets
- Sensitivity analysis can be an alternative to pooling
- Multiple imputation fixes inferences, but still has the SDD assumption!