3.08 Missing Data

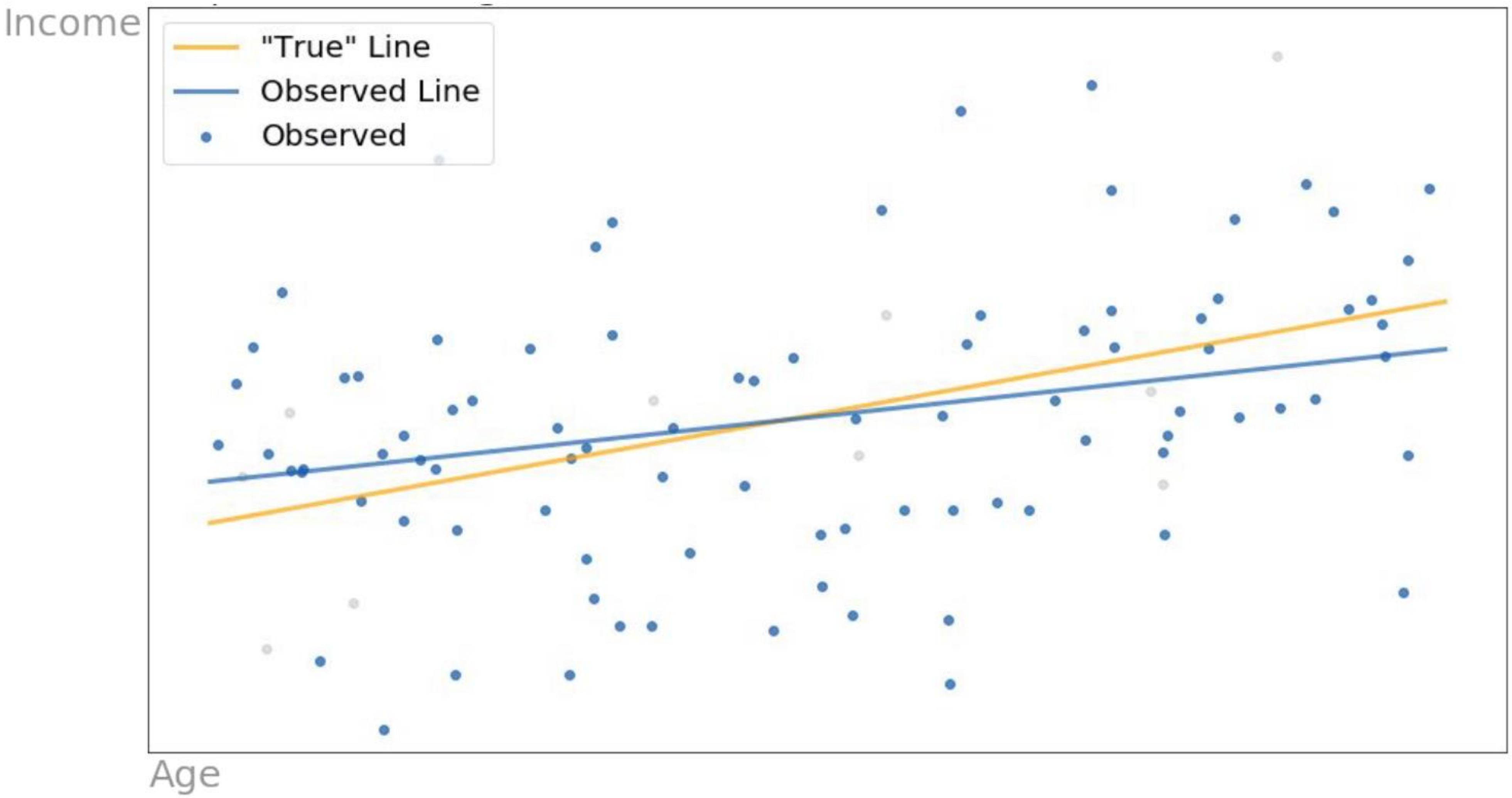
How big of a problem is missing data?

This is a difficult question to answer.

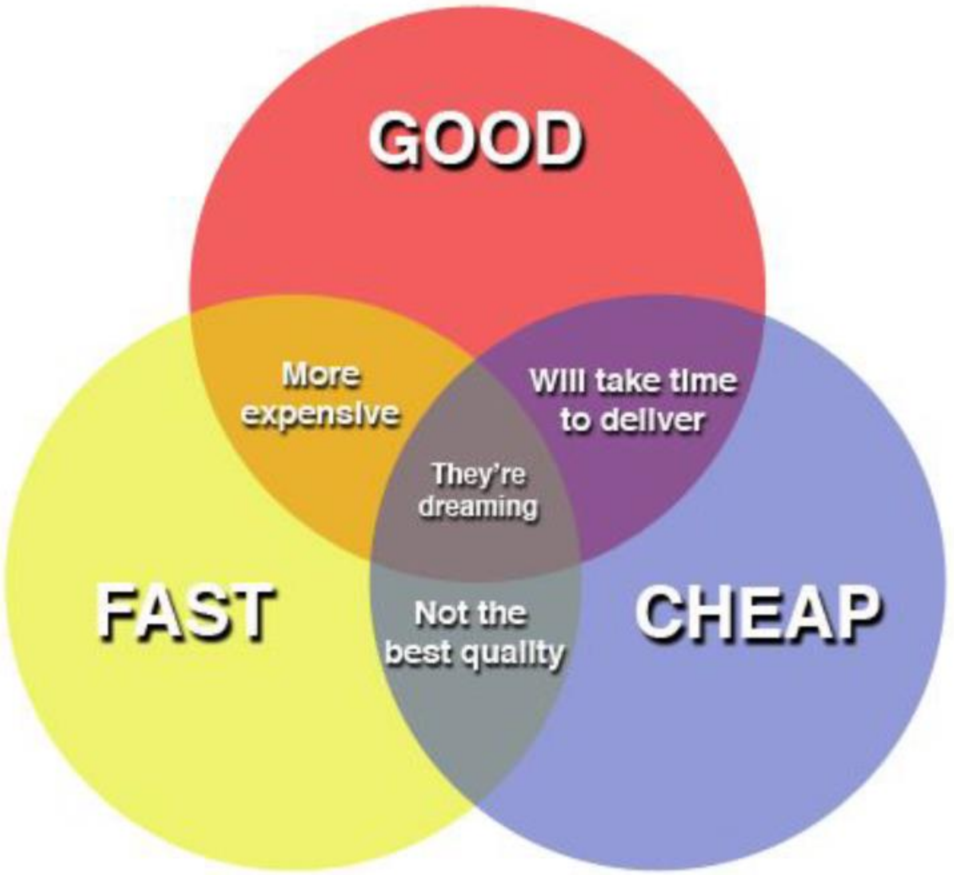
Practically, we can only see what we observe.

We can use simulated data to help answer this question.

Let’s say we only have 10% of our data missing



What is a realistic approach for us?



http://www.pyragraph.com/2013/05/good-fast-cheap-you-can-only-pick-two/

What is a realistic approach for us?

Fast and Cheap Analysis: Drop all missing values or single imputation.

Good and Cheap Analysis: Proper imputation or pattern submodel method.

Good and Fast Analysis: Gather data in a complete manner.

http://www.pyragraph.com/2013/05/good-fast-cheap-you-can-only-pick-two/

Strategies for doing data science with missing data

1. Avoid missing data
2. Ignore missing data
3. Account for missing data
    a. Unit missingness
    b. Item missingness

Strategy 1: Avoid Missing Data

It’s often more expensive up front but **cheaper in the long run** to **avoid missing data** than to make guesses about how to best handle our missing data.

- Decrease burden on your respondent.
- Change method of data collection.
- Improve accessibility.
- Change timing of your survey.
- Minimize length of questionnaire.
- Consider content of your survey.

Strategy 2: Ignore Missing Data

We **assume** that our observed data is **similar to our missing data**.

One general, ***very rough*** *guideline* is that we may be OK ignoring missing data if less than 5% of our data is missing.

- If we're doing supervised learning and we're missing a lot of our Y variable, this may be inadvisable.
- If we're missing a lot from meaningful variables, this may be inadvisable.

Strategy 3: Account for Missing Data

There's a naive belief that we can just plug in the gaps in our data.

- This is known as **imputation**.
- We have to do this in a specific way, or we're just making up data.

In most cases, we aren't "fixing" data. We're just learning how to cope with it!

Unit vs. Item Missingness

Unit missingness has all values missing from an observation.

- Index 3.

Item missingness is where some, but not all, values are missing from an observation.

- Indices 1, 2, and 10,000.

Index	Age	Sex	Income
1	NA	M	NA
2	39	NA	75000
3	NA	NA	NA
4	28	F	50000
...	...	...	...
10000	18	F	NA

Types of Missingness

1. Missing Completely at Random (MCAR)

2. Missing at Random (MAR)

3. Not Missing at Random (NMAR)

Scenario 1: Missing Completely at Random (MCAR)

I'm a grad student in a lab. While pipetting, I reach for my pen but accidentally knock a Petri dish off of the desk. From this Petri dish, I lose the data that I otherwise would have collected.

This is called **missing completely at random**.

- The data of interest is not systematically different between missing and observed.

bacteria on day 1	bacteria on day 2
10mm	15mm
12mm	12mm
9mm	11mm
10mm	11mm
15mm	19mm
13mm	15mm
11mm	~~16mm~~

Scenario 2: Missing at Random (MAR)

I work for the Department of Transportation. A sensor on the Pennsylvania Turnpike broke and did not gather information between 7:00am and 10:00am.

This is called **missing at random**.

- **Conditional on data we have observed**, the data of interest is not systematically different between missing and observed.
- Whether or not a data point is missing is dependent on observed data.

time	number of vehicles
4:00	206
5:00	519
6:00	934
7:00	~~1,650~~
8:00	~~1,921~~
9:00	~~1,010~~
10:00	889

Scenario 3: Not Missing at Random (NMAR)

I administer a survey with a question about income. Those who have lower incomes are less likely to reply to the income question.

This is called **not missing at random**.

- The data of interest are systematically different for missing and observed.
- Whether or not an observation is missing depends on the value of the unobserved data itself!

id	income
A	~~48,000~~
B	~~35,000~~
C	105,000
D	~~62,000~~
E	80,000
F	50,000
G	75,000

Missing Values Workflow

- Evaluate size of missing data
- Decide if:
  - Is it worth your time to try to address it?
  - Is it reasonable to attempt deductive imputation?
- If deductive imputation is not feasible, you may use:
  - central tendency imputation
  - regression imputation
  - nearest neighbour imputations

Warning 1

If your goal is just to have a “complete” data set for further analysis, **be very careful**!

- After you construct this dataset, **nobody will know the difference between observed and imputed data**.
- At the end of the day, you could be **making up data**.

Warning 2

There are **`sklearn`** methods (**`SingleImputer`**, **`IterativeImputer`**) that can be used.

However, **`IterativeImputer`** is currently experimental!

Proceed with caution.

What is a realistic approach for us?

Fast and Cheap Analysis: Drop all missing values or single imputation.

Good and Cheap Analysis: Proper imputation or pattern submodel method.

Good and Fast Analysis: Gather data in a complete manner.

http://www.pyragraph.com/2013/05/good-fast-cheap-you-can-only-pick-two/