Identification and Missing Data: A Most Excellent Read

Suh Young Choi

Mentors: Eric Morenz & Yiqun Chen

Why am I keeping you here for the next 10 minutes???

- Recap of topics in DRP
- Overview of my project
- Thinking about the future

http://knowledgequest.aasl.org/wp-content/uploads/2019/05/GoodReads-logo.jpg

Overview of DRP topics

M*ssing d*ta? N*ver he*rd of *t!

Missing Data

- It's a thing!
- How to deal with missing data?

  It depends on the source of missingness.

  - Missing completely at random (MCAR)
  - Missing at random (MAR)
  - Missing not at random (MNAR)

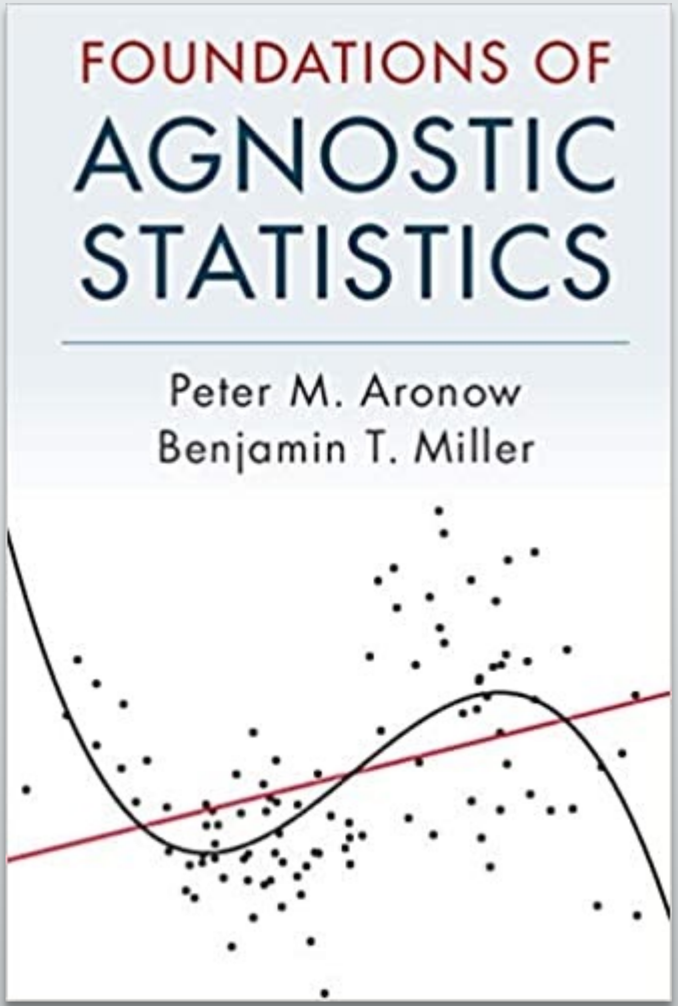


image credit: GoodReads

Review on the source of missingness

- **MCAR**: There is no relationship between the missingness of the data and any values, observed or missing.

- **MAR**: There is a systematic relationship between the propensity of missing values and the observed data, but not the missing data.

- **MNAR**: There is a relationship between the propensity of a value to be missing and its values.

Fantastic Projects and Where to Find Them

Answer: my working directory

Questions of interest

- Learned several methods for dealing with missing data

1. Simulate missingness in **a variable** from a complete dataset
2. Apply these methods to obtain **a point estimate** of **the variable**
3. Interpret and compare estimates from different methods, including one from the original data

- This will help us understand the
  relationship between **the variable of interest** with
  other variables as an example,
  but not limited to this.
  It can be extended to other cases!

Questions of interest

- Learned several methods for dealing with missing data

1. Simulate missingness in *number of ratings* from a complete dataset
2. Apply these methods to obtain sample mean of *number of ratings*
3. Interpret and compare estimates from different methods, including one from the original data

- This will help us understand the relationship between *number of ratings* v. *pages* as an example, but not limited to this.

  It can be extended to other cases!

Data for this project

- source: https://www.kaggle.com/jealousleopard/goodreadsbooks
- data created 2019-06-14, last updated 2020-03-19
- 10294 books

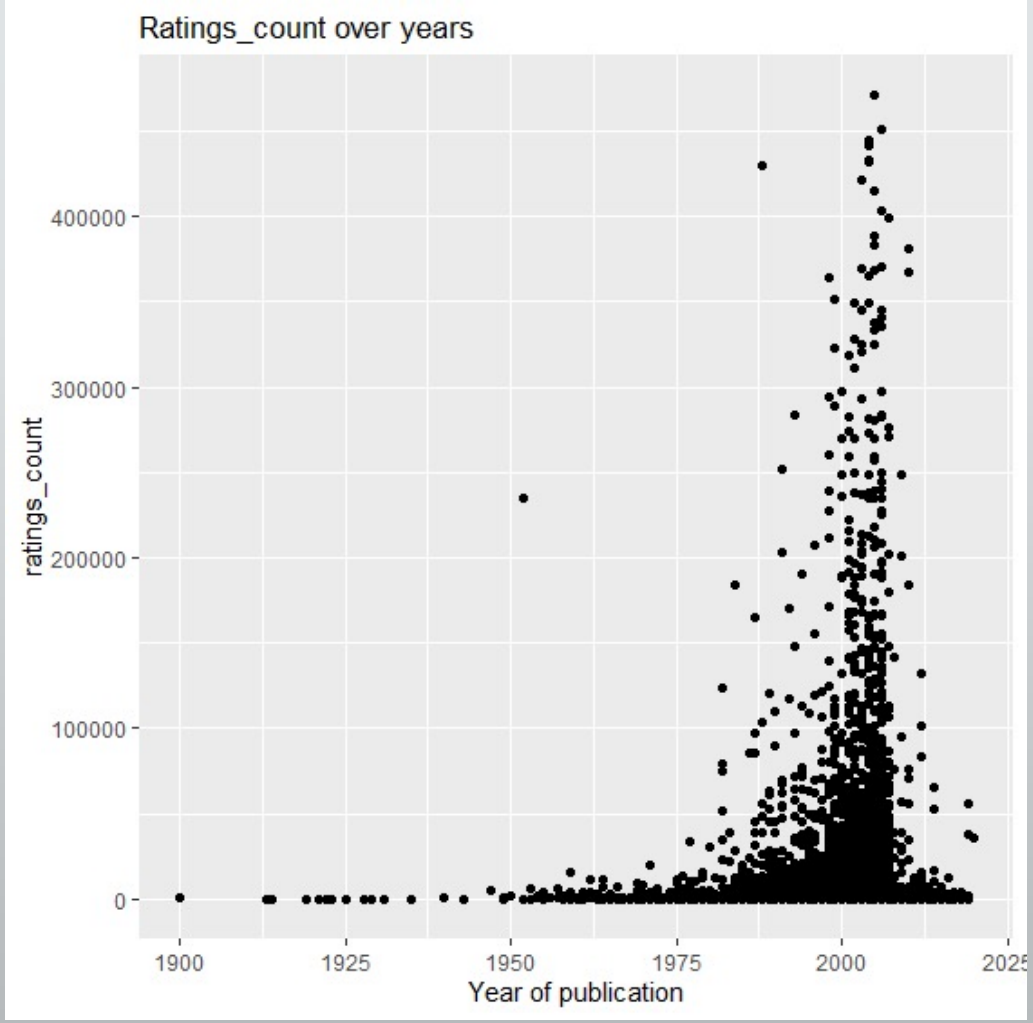
variable	descriptions	remark
title	The name under which the book was published.	
authors	Names of the authors of the book.	
num_pages	Number of pages the book contains.	
ratings_count	Total number of ratings the book received.	variable of interest
text_reviews_count	Total number of written text reviews the book received.	
pub.yearonly	Year when the book was published.	1900 – 2020

Missingness comes in!

- MAR: Replace ratings_count by NA if publication year < 1990

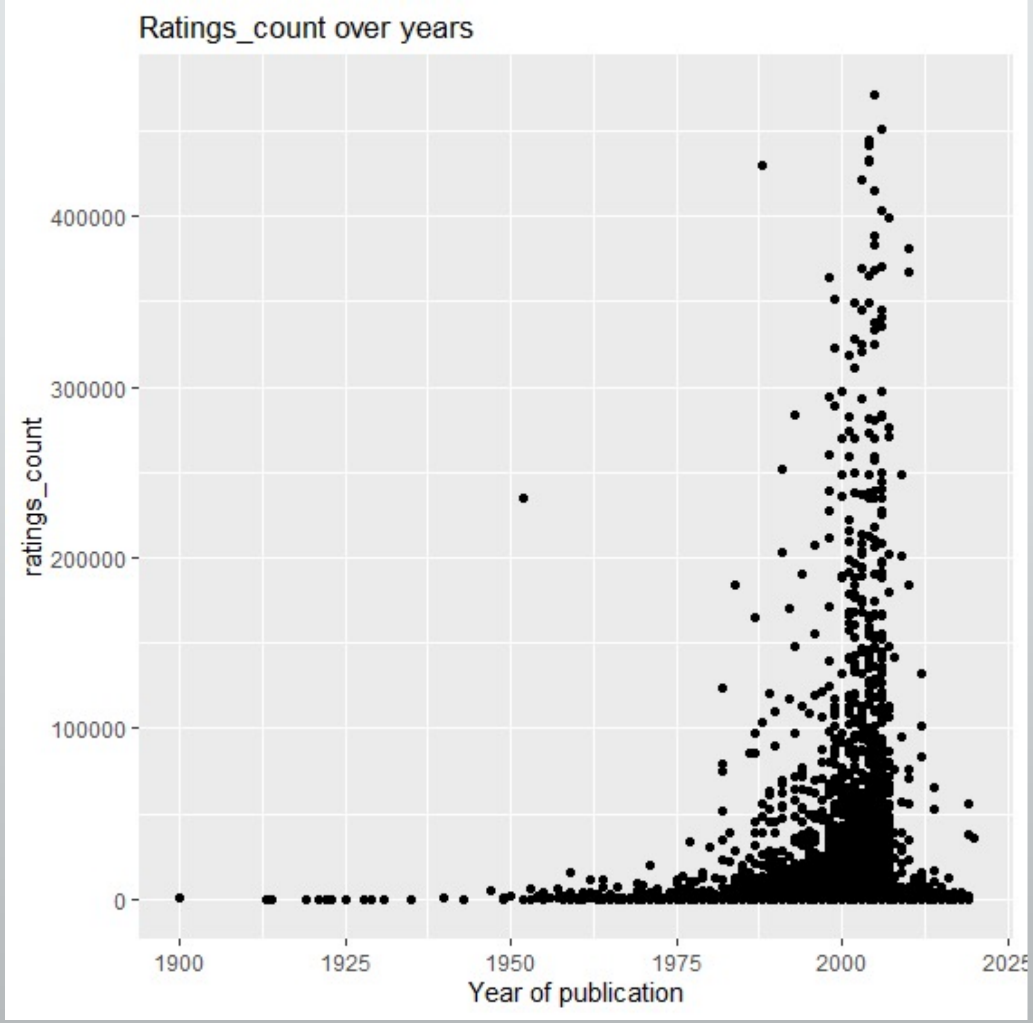
Missingness comes in!

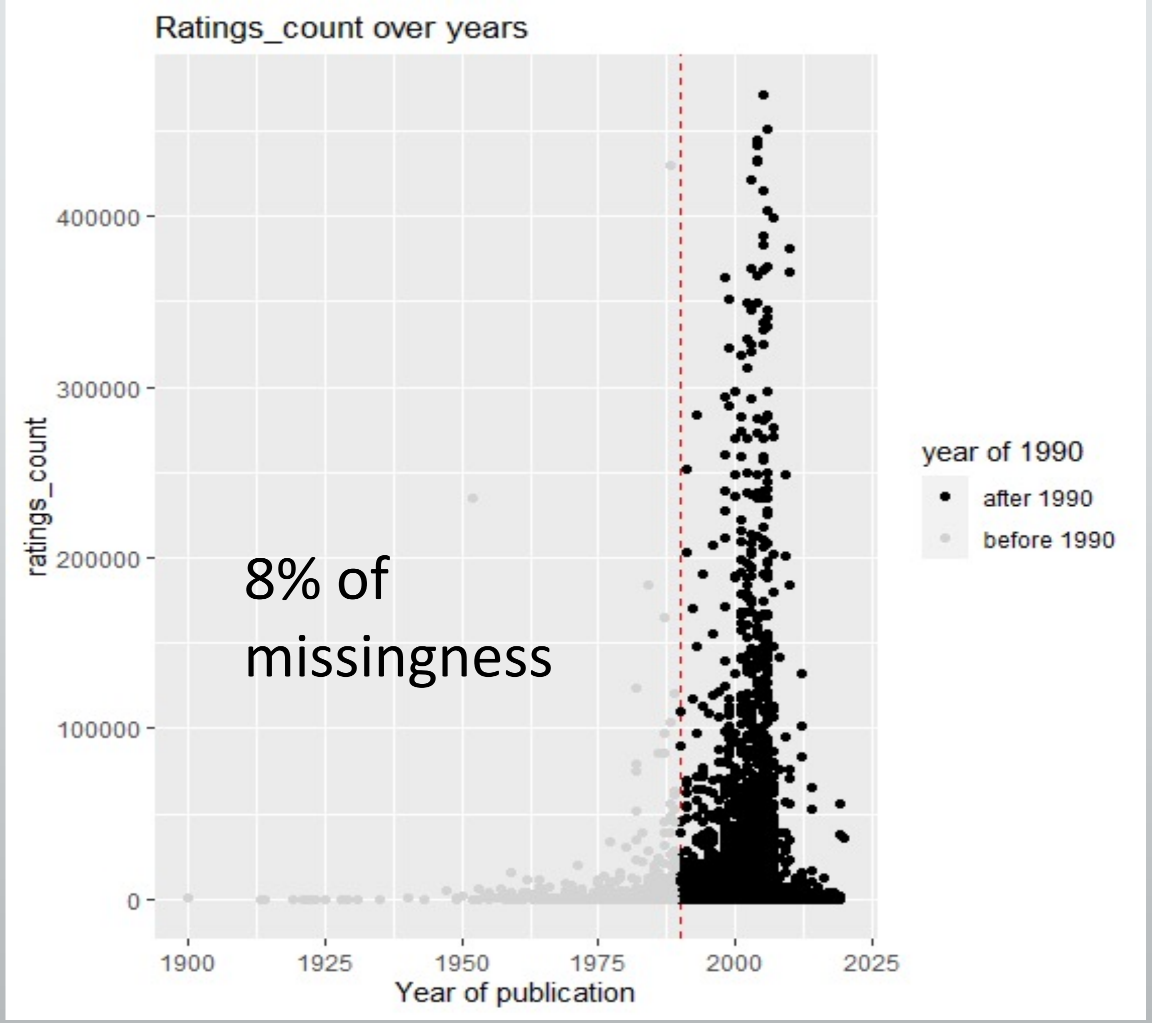
- MAR: Replace ratings_count by NA if publication year < 1990



Missingness comes in!

- MAR: Replace ratings_count by NA if publication year < 1990





8% of missingness

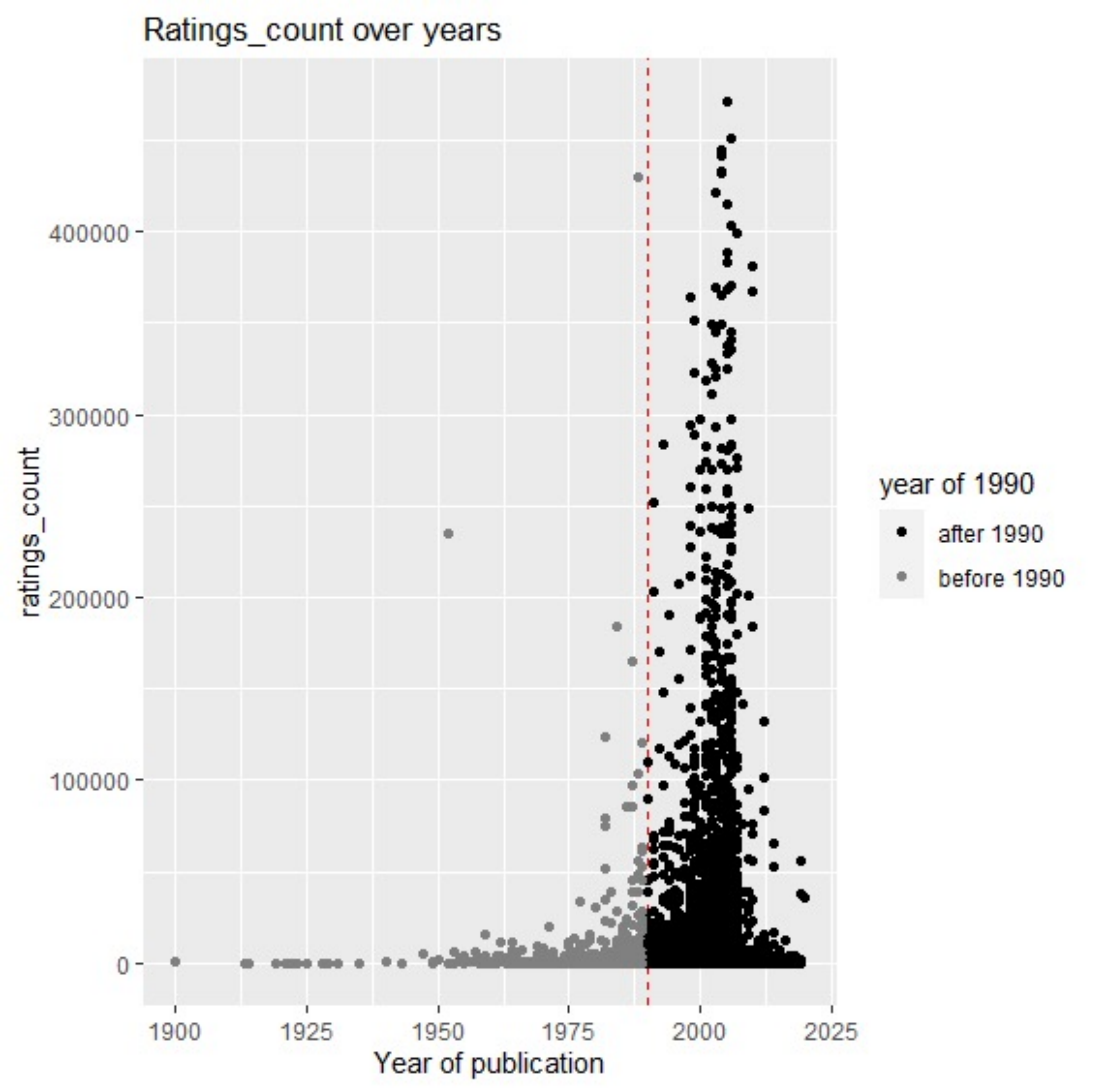
Estimators & estimates

(*) imputation replaces NA by mean estimate
(**) imputation replaces NA by covariate value, proportionally (X: 1 in 10's, 0 otherwise)

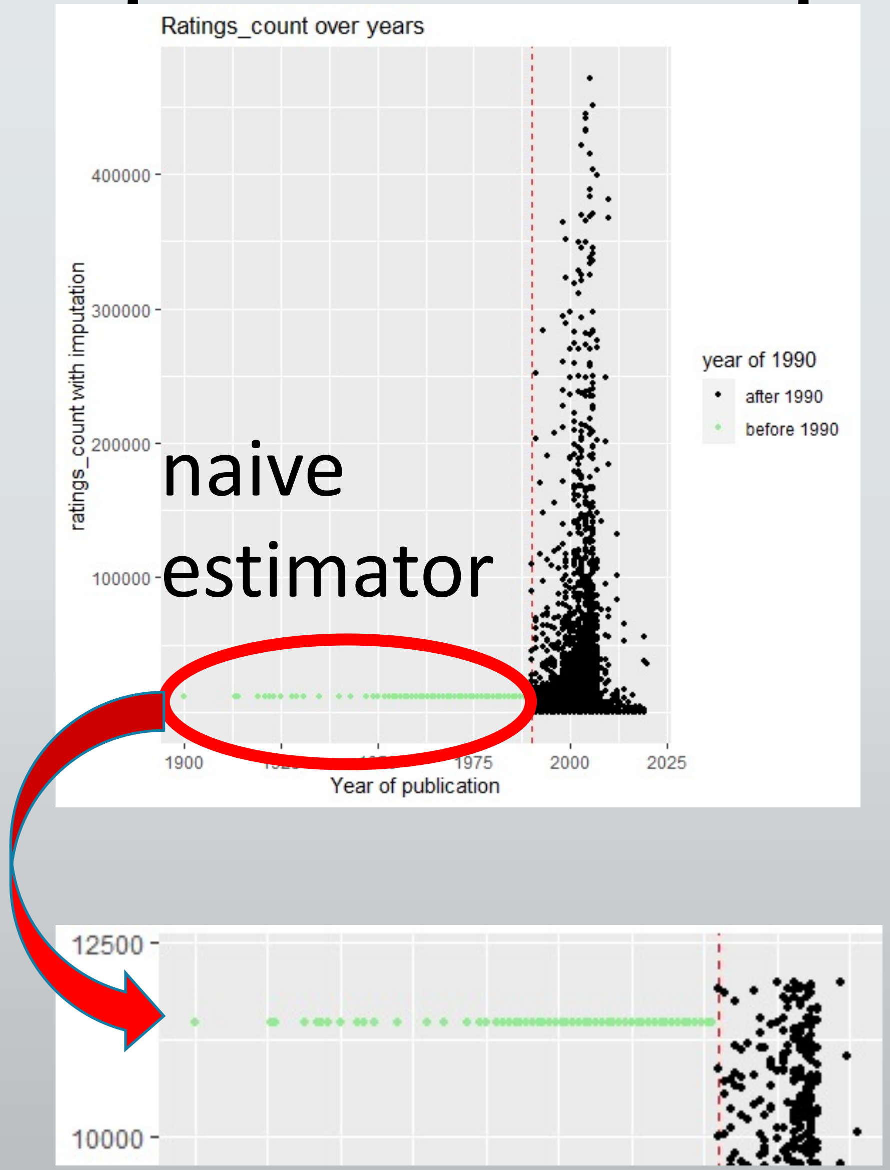
- Y: variable of interest, "*ratings_count*"
- X: binary covariate
- R: indicator if Y is observed (1) or missing (0)

Estimators	Estimates of mean	Remark
Mean from the original data (full set)	10861.9	(true mean value)
Excluding missing values (naïve estimator)	11459.91	mean of observed values; proper for MCAR (*)
Plug-in estimator	11462.43	post-stratification estimator for missing data (**)
Regression estimator	11462.37	use OLS regression to approximate conditional expectation function (CEF) of Y given X
Inverse Probability-Weighted (IPW) estimator	11462.37	inverse-weighted using the response propensity function (RPF), P(R=1|X): useful when strong ignorability holds.
Doubly Robust (DR) estimator	11462.37	if either the approximation of CEF or approximation of RPF are exactly correct

comparison on imputation

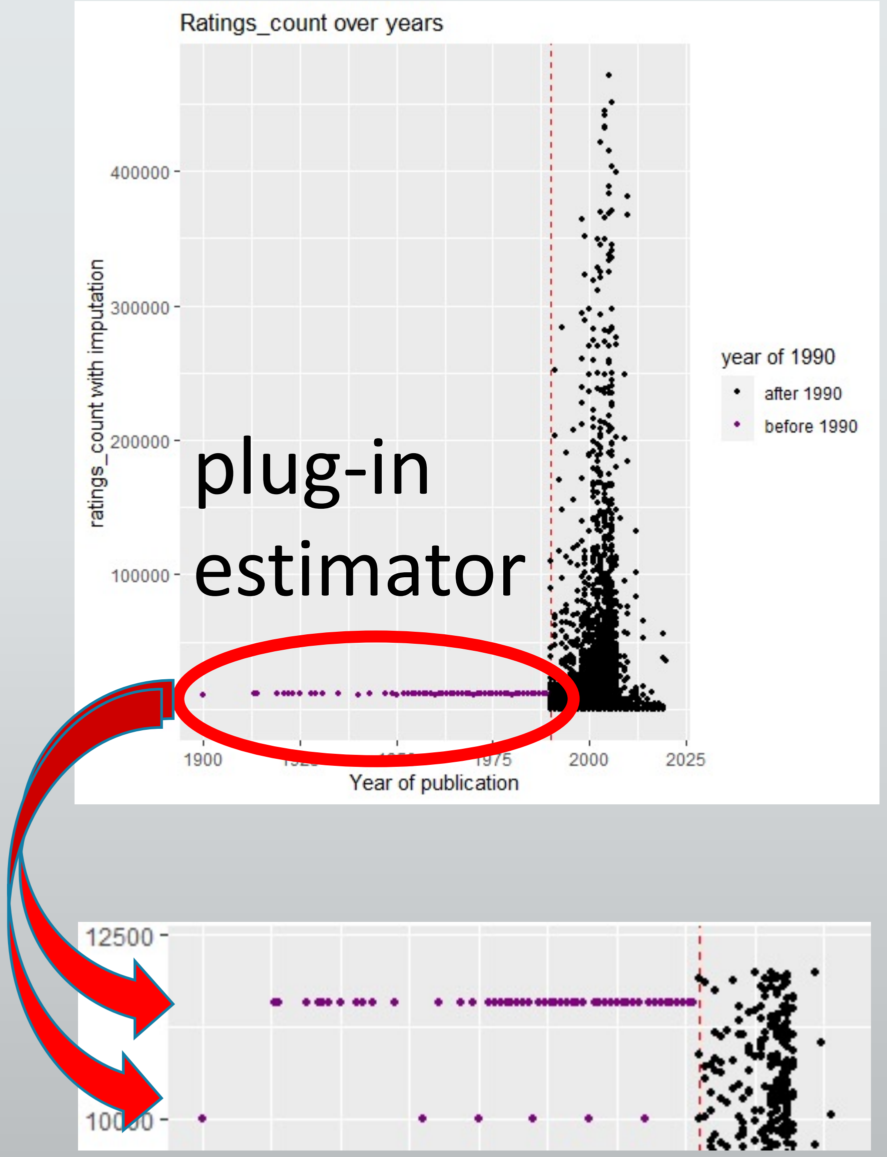


comparison on imputation



naive estimator

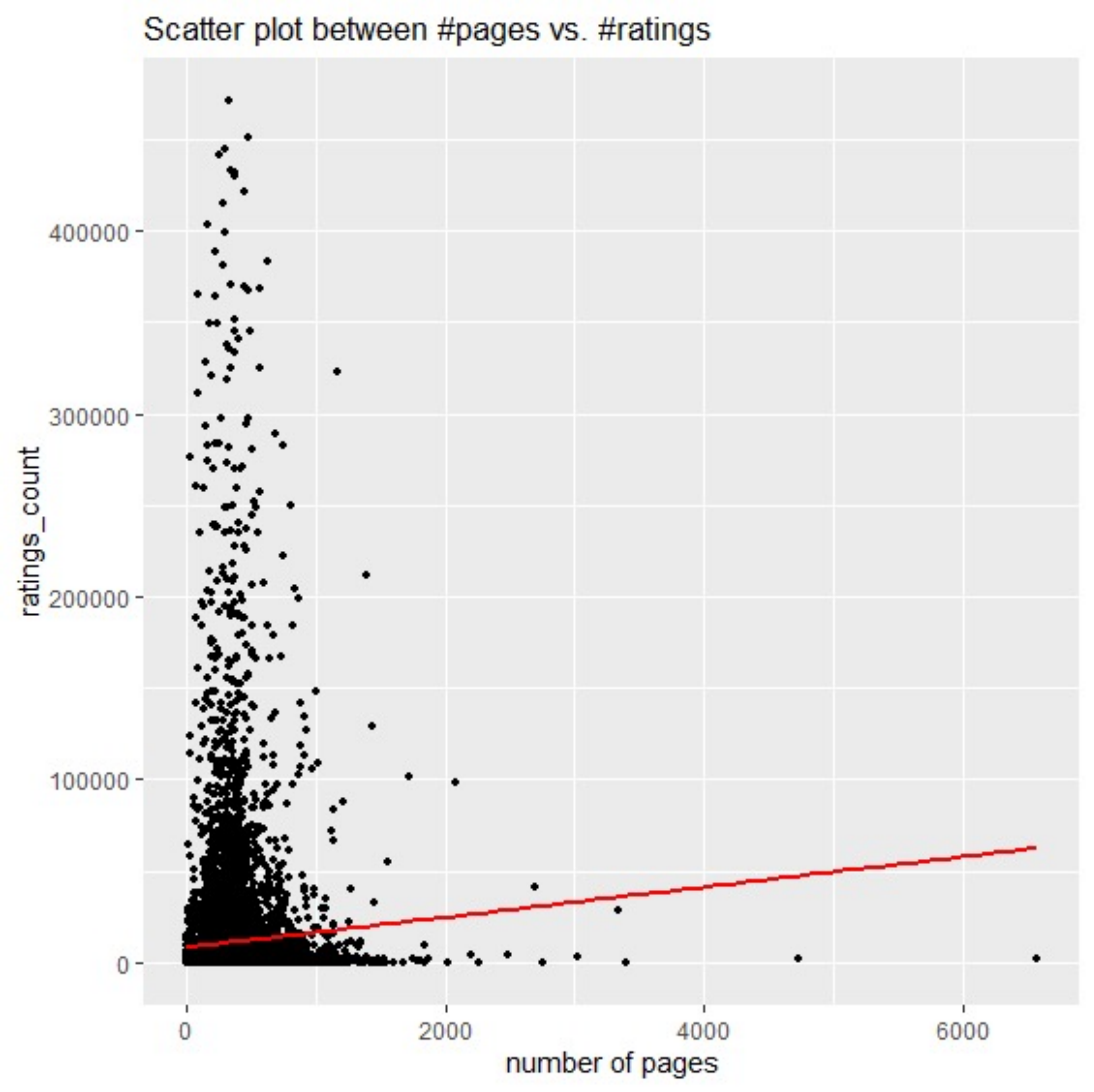
imputed by a constant



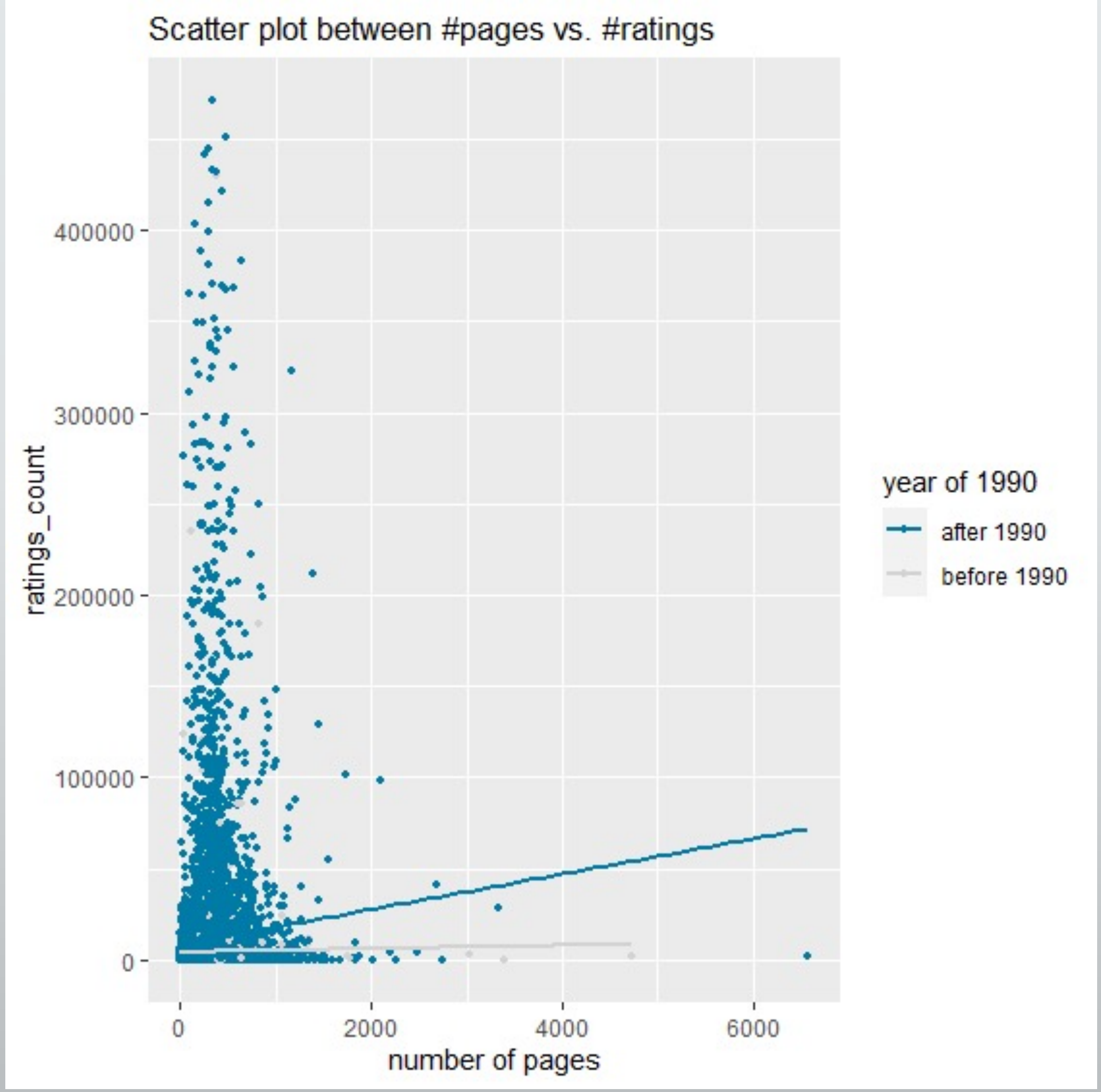
plug-in estimator

imputed by constants depending on X

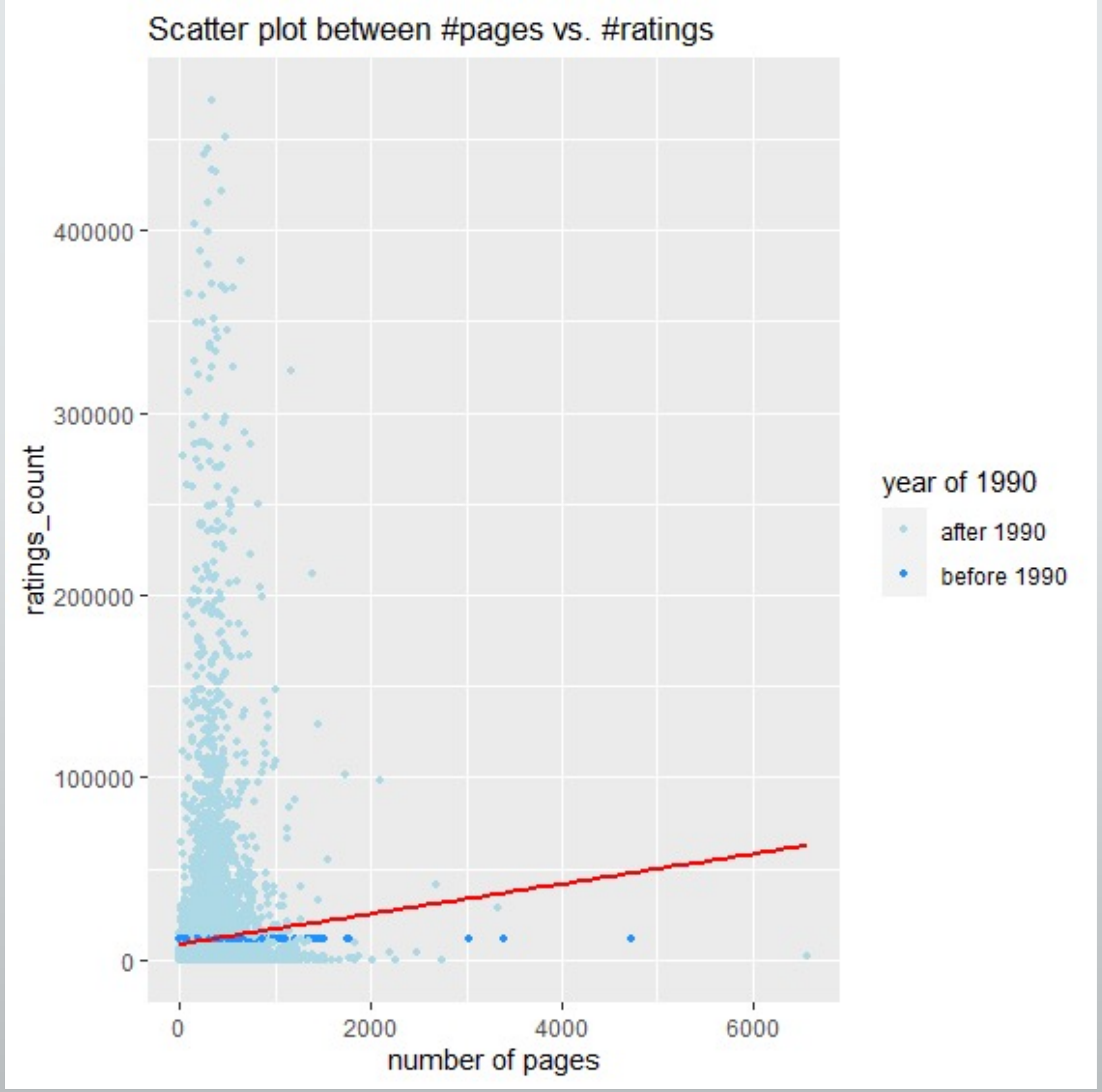
full data



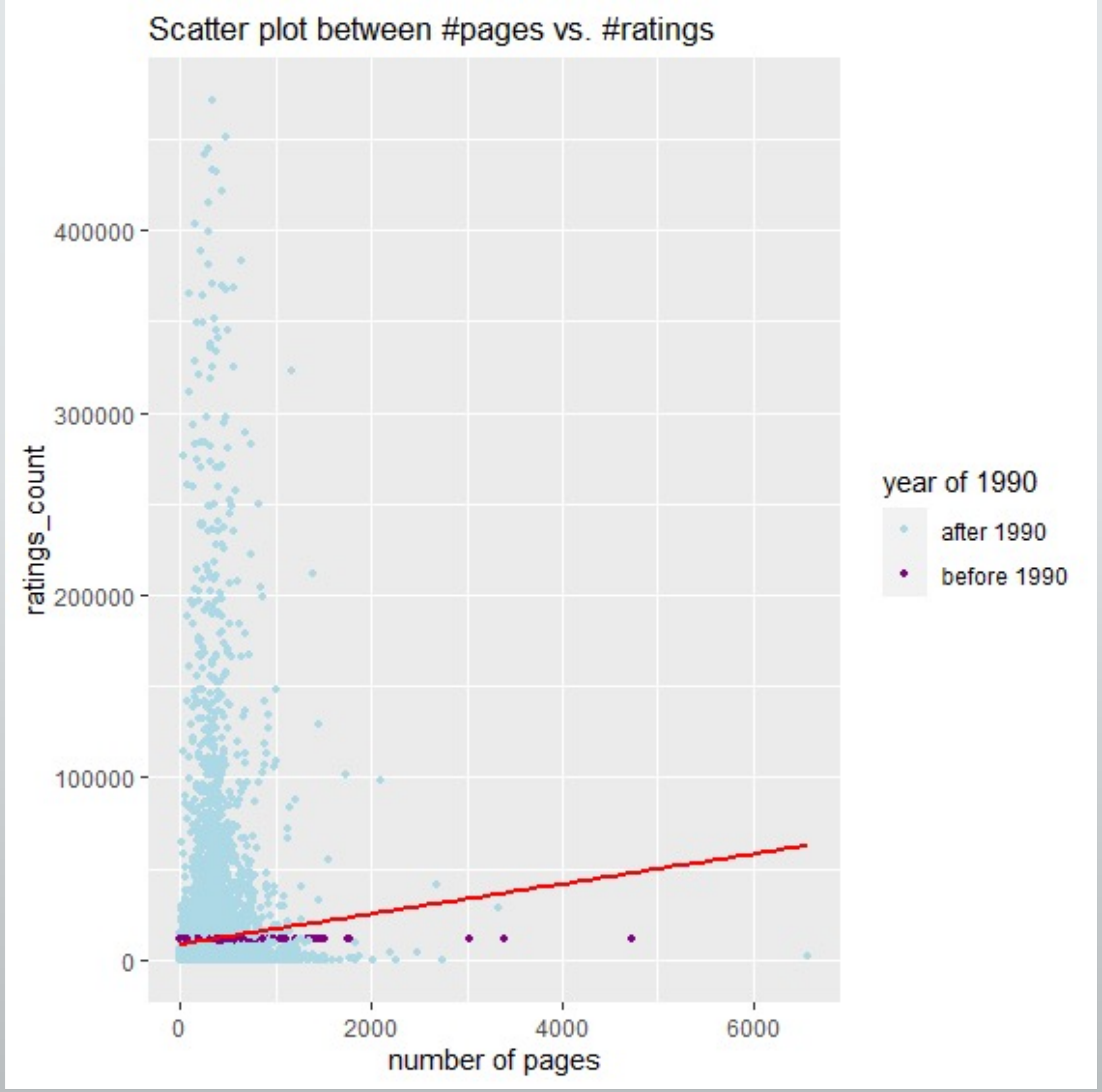
with MAR on ratings_count



imputed by naive estimator



imputed by plug-in estimator



Concluding Remarks

- The naive estimator is for MCAR, and the simplest approach.
- There are more than one way to handle MAR, and the choice depends on the covariates.
- It also helps us to understand the source of missingness.

Acknowledgments

Many thanks to Yiqun, Eric, and Thomas for a great quarter!

Questions?