Missing Data

and how to deal with them...

Andreas Wittmann

2021/07/09 (updated: 2021-07-08)

Missing data

Can arise for many reasons:

- Non-Response e.g. in surveys
- Lost data due to machine or human mistakes
- Bug issues in non-mandatory fields
- join, merge
- Different variable per source
- Different number of categories per source
- …

The problem

```r
x <- c(10, NA, 20, 30, 40, 20)

mean(x)
```

```
[1] NA
```

```r
randomForest(Ozone ~ . , data=airquality)
```

```
Error in na.fail.default(structure(list(Ozone = c(41L, 36L, 12L, 18
```

The problem

- Many AI/ML/Data science methods are developed for complete data
- Using only the complete cases for the analysis can lead to dramatic information loss
- Inappropriate approach imposes noise or bias on data
- Can lead to incorrect conclusions due to absense of relevant information
- The quality of statistical analysis can be only as good as the quality of the data

Terminology

- **Full / complete data** $Z = (Z^{\text{obs}}, Z^{\text{mis}})$
- **Observed data** Z^{obs}
- **Unobserved / missing data** Z^{mis}
- **Complete cases** subset of rows without missing values
- Given $n \times p$ data matrix Z, which can contain missing data
- $Z = (Y, X)$, i.e. Y matrix dependent and X matrix independent variables
- Indicator matrix R build from Z as

$$R_{ij} = \begin{cases} 1 & \text{if } Z_{ij} \text{ obs} \\ 0 & \text{if } Z_{ij} \text{ mis} \end{cases} \qquad \text{for } i = 1, \ldots, n \text{ and } j = 1, \ldots, p.$$

Types of missingness

Missing completely at random (MCAR)

Probability of missingness is completely independent from observed and unobserved/missing values:

$$P(r_i \mid z_i) = P(r_i \mid z_i^{\text{obs}}, z_i^{\text{mis}}) = P(r_i), \quad \text{for } i = 1, \dots, n,$$

z_i^{obs} observed, z_i^{mis} missing values from the i-th row z_i of the data matrix Z

- No particular reason that the data is missing
- Often an unrealistic assumption
- **Example:** Weighing scale that ran out of batteries

Missing at random (MAR)

Probability of missingness of values is only dependent of the observed values z_i^{obs}

$$P(r_i \mid z_i) = P(r_i \mid z_i^{\mathrm{obs}}, z_i^{\mathrm{mis}}) = P(r_i \mid z_i^{\mathrm{obs}}), \quad \text{for } i = 1, \ldots, n.$$

- More realistic than MCAR
- Modern missing data methods generally start from the MAR assumption
- **Example:** Weighing scale may produce more missing data when placed on a soft surface and type of surface is known

Missing not at random (MNAR)

Probability of missingness of values is dependent of the observed z_i^{obs} and unobserved values z_i^{mis}

$$P(r_i \mid z_i) = P(r_i \mid z_i^{\text{obs}}, z_i^{\text{mis}}), \quad \text{for } i = 1, \ldots, n.$$

- Cause of missingness it not known
- We cannot draw any conclusion from observed data
- **Example:** Weighing scale mechanism may wear out over time, but time is not part of the dataset

How to deal with missingness

Strategies to deal with missing data

- Prevention - impossible for ex-post analyses
- Dropping missing values
- Imputation techniques
  - Single imputation
  - Multiple imputation

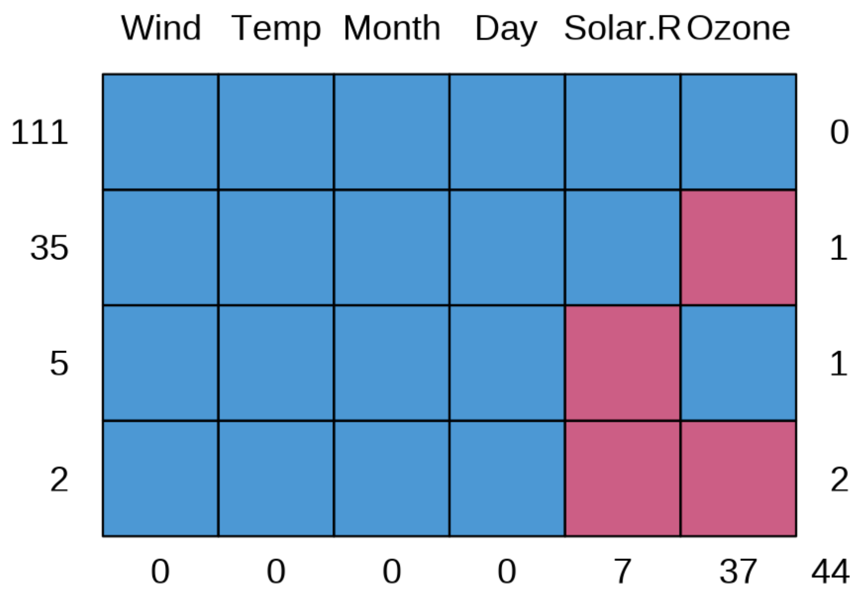
Look at the data

Airquality Dataset

- Daily air quality measurements in New York, May to September 1973.
- Daily readings of the following air quality values for May 1, 1973 (a Tuesday) to September 30, 1973.
  - **Ozone:** Mean ozone in parts per billion from 1300 to 1500 hours at Roosevelt Island
  - **Solar.R:** Solar radiation in Langleys in the frequency band 4000–7700 Angstroms from 0800 to 1200 hours at Central Park
  - **Wind:** Average wind speed in miles per hour at 0700 and 1000 hours at LaGuardia Airport
  - **Temp:** Maximum daily temperature in degrees Fahrenheit at La Guardia Airport.

Source: The data were obtained from the New York State Department of Conservation (ozone data) and the National Weather Service (meteorological data).

Missing data pattern



Dropping (ignoring) missing values

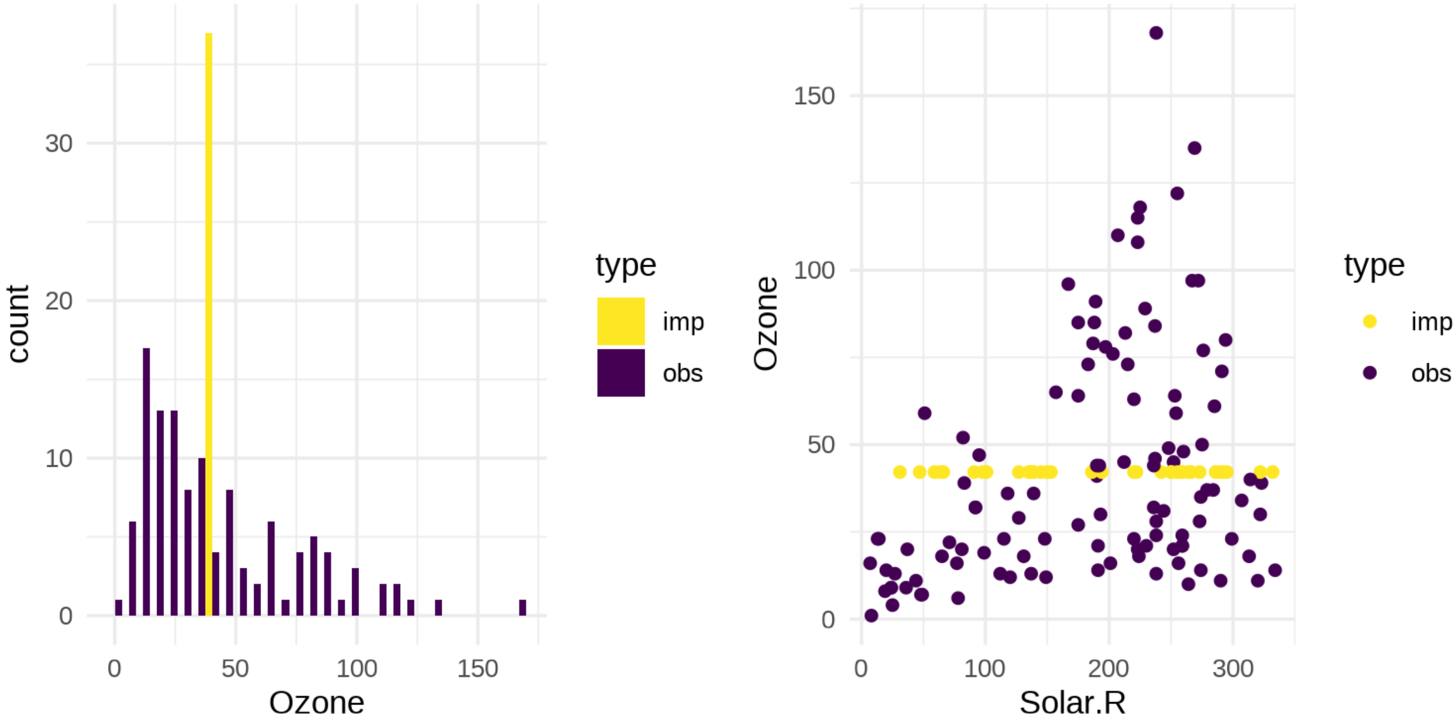
Listwise deletion

- Only the complete cases are analyzed
- Advantages:
  - Simple - Often the default way of handling incomplete data
  - Under MCAR: unbiased estimates of means, variances and regression weights
  - Schafer and Graham (2002): *If a missing data problem can be resolved by discarding only a small part of the sample, then the method can be quite effective.*
- Disadvantages:
  - Loss of information dependent on the fraction of missing data
  - Larger standard errors
  - Under MAR: biased, even for simple statistics like the mean

Mean/Median imputation

- Missing values are replaced by
    - The mean value for quantitative variables
    - The most frequently occurring category for qualitative variables
- Imputed value is an estimate, thus there is uncertainty about its true value
- Uncertainty is measued by its standard error
- Too small standard errors

Mean/Median imputation

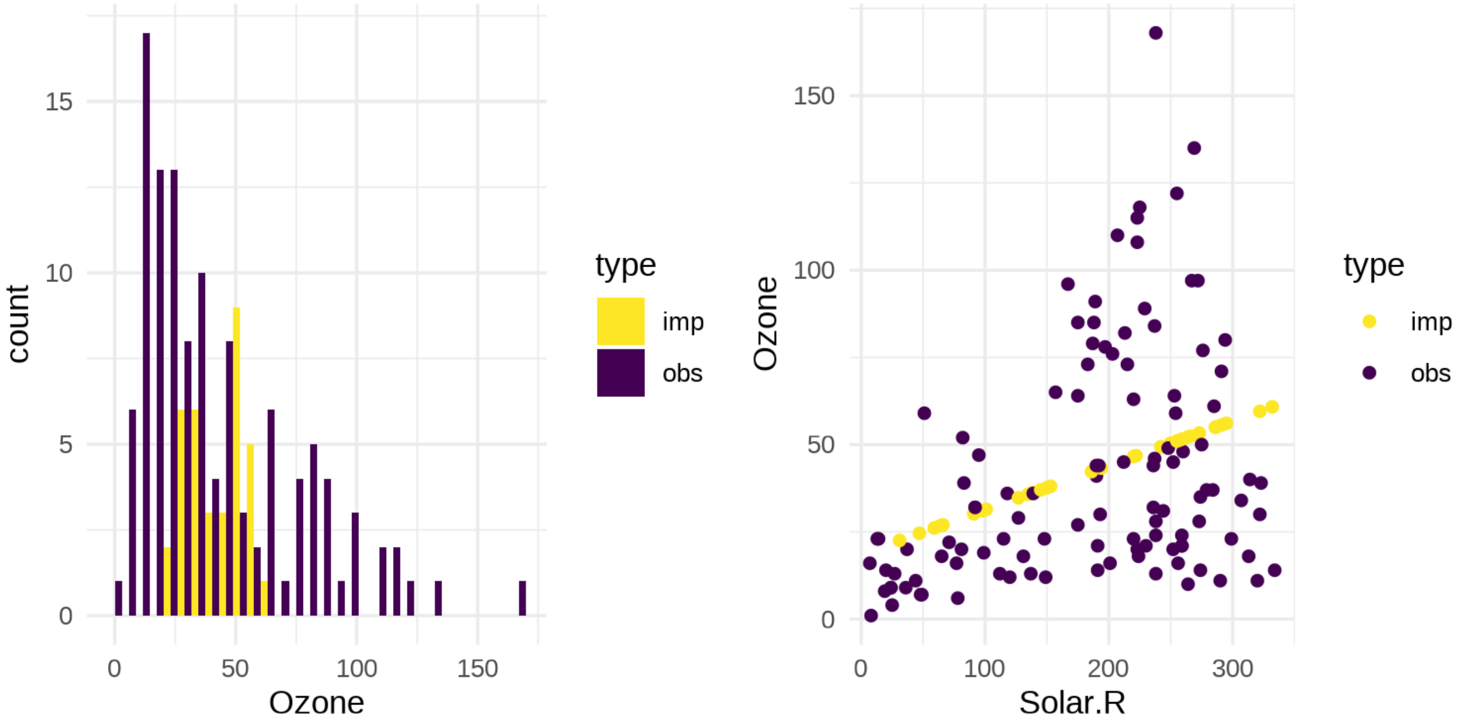


Regression Imputation

- Regression imputation incorporates knowledge of other variables
- The first step involves building a model from the observed data
- Calculate predictions for the incomplete cases under the fitted model

$$\mathrm{Ozone} = \alpha + \beta_1(\mathrm{Solar.R}) + \epsilon$$

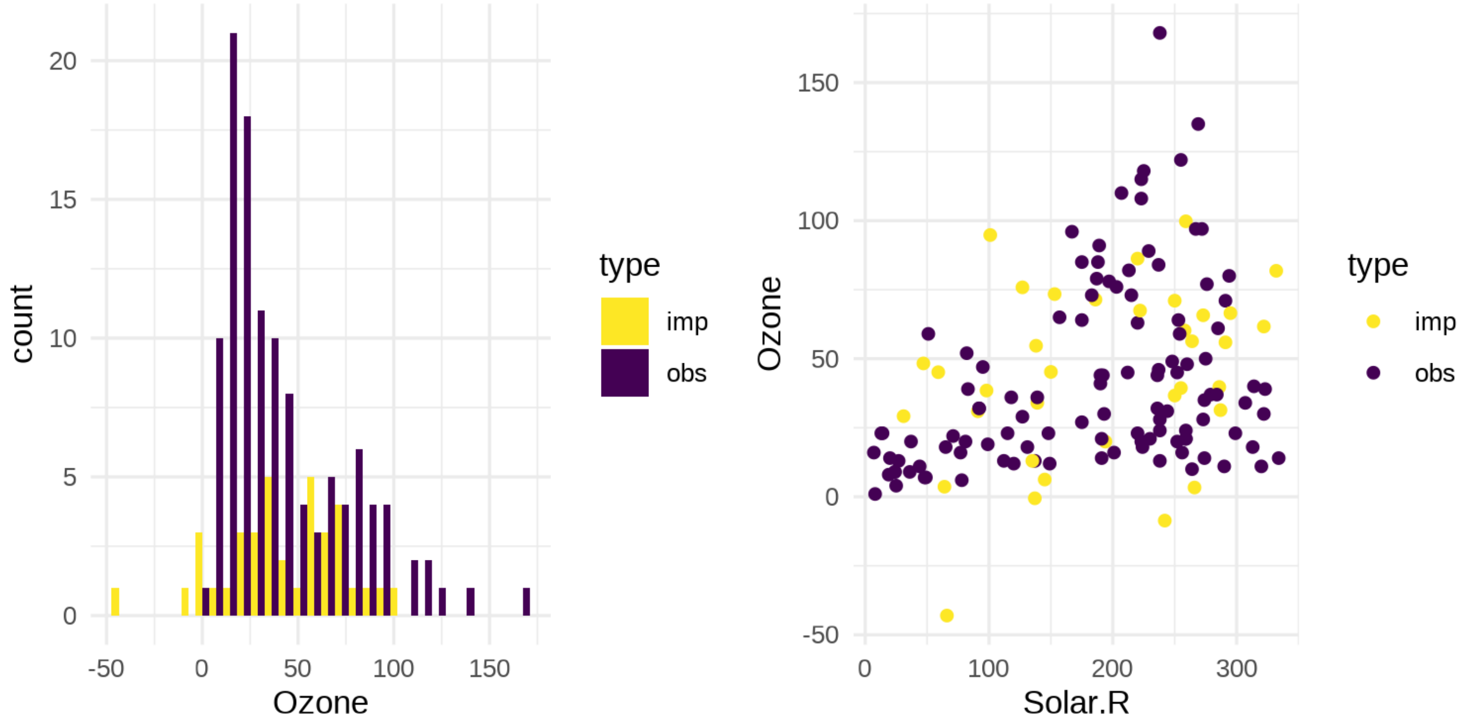
Regression Imputation



Stochastic Regression Imputation

- Regression imputation disadvantage:
  - Fitted model is used without error terms
  - Imputed results are too close to the regression line
  - Biased correlations, reduced the variance of the data
- Stochastic regression adds an error term when imputing the values

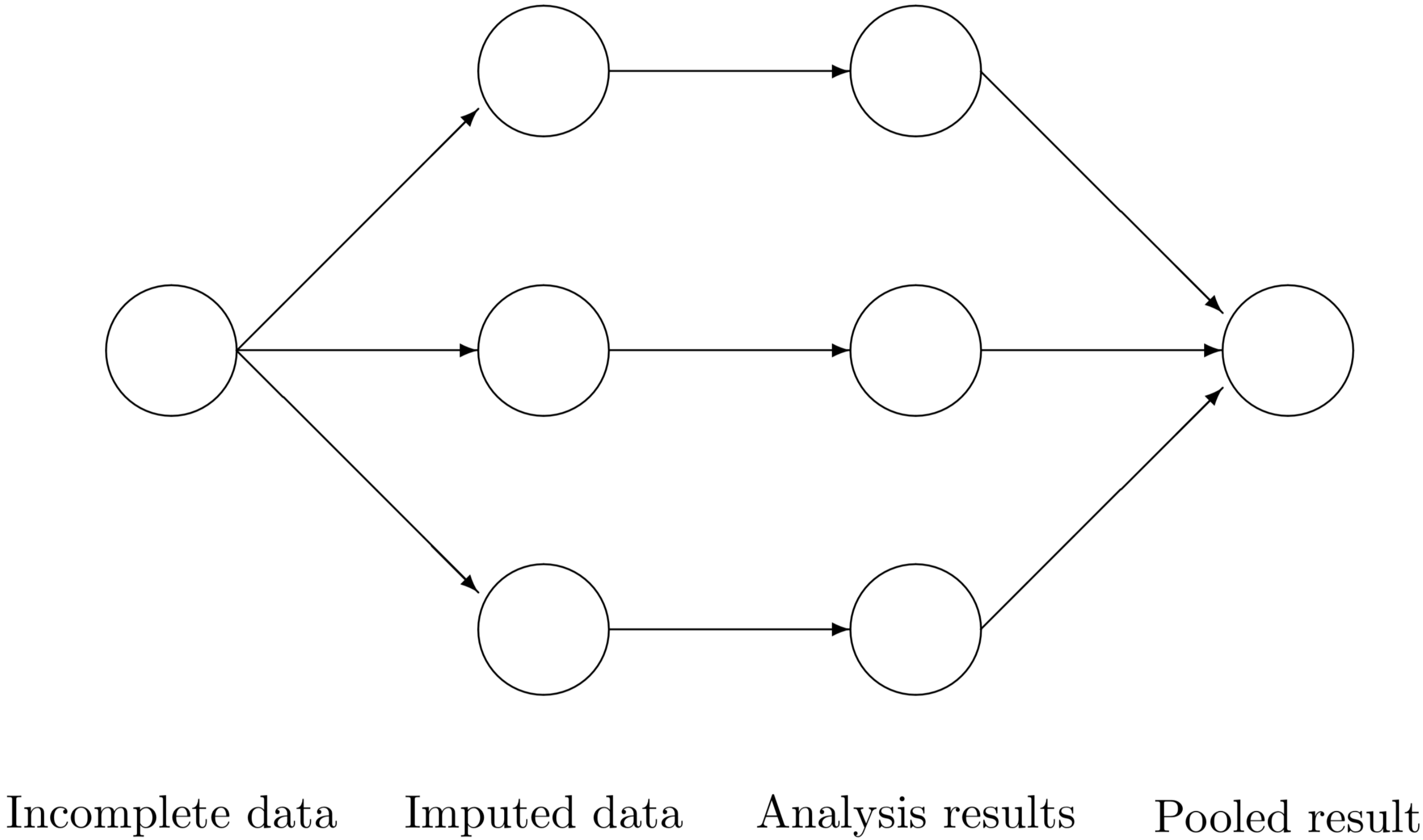
Stochastic Regression Imputation



Multiple imputation

- Accounts for uncertainty by creating multiple imputed version of data
- Bootstrapping (subselection of the data, do the imputation, …)
- Generative models (draw samples from the estimated distribution)
- MICE (multivariate imputation by chained equations): Missing values are replaced by chained regression, where m complete datasets are generated (Raghunathan, Lepkowski, Van Hoewyk, et al. (2001))

Multiple imputation



Source: https://stefvanbuuren.name/fimd/sec-nutshell.html

Multiple imputation (vanBuuren (2018))

1. Specify an imputation model $P(Y_j^{\text{mis}}|Y_j^{\text{obs}}, Y_{-j}, R)$ for variable Y_j with $j = 1, \ldots, p$.
2. For each j, fill in starting imputations $\dot{Y}_j^0$ by random draws from Y_j^{obs}.
3. Repeat for $t = 1, \ldots, m$.
4. Repeat for $j = 1, \ldots, p$.
5. Define $\dot{Y}_{-j}^t = (\dot{Y}_1^t, \ldots, \dot{Y}_{j-1}^t, \dot{Y}_{j+1}^{t-1}, \ldots, \dot{Y}_p^{t-1})$ as the currently complete data except Y_j.
6. Draw $\dot{\phi}_j^t \sim P(\phi_j^t|Y_j^{\text{obs}}, \dot{Y}_{-j}^t, R)$.
7. Draw imputations $\dot{Y}_j^t \sim P(Y_j^{\text{mis}}|Y_j^{\text{obs}}, \dot{Y}_{-j}^t, R, \dot{\phi}_j^t)$.
8. End repeat j.
9. End repeat t.

Multiple imputation in detail...

1. Start

Ozone	Solar.R	Wind	Temp	Month	Day
NA	194	8.6	69	5	10
7	NA	6.9	74	5	11
16	256	9.7	69	5	12

2. Mean Imputation

Ozone	Solar.R	Wind	Temp	Month	Day
11.5	194	8.6	69	5	10
7.0	225	6.9	74	5	11
16.0	256	9.7	69	5	12

Multiple imputation in detail...

3. Set Ozone to NA's / Regression on complete cases

Ozone	Solar.R	Wind	Temp	Month	Day
NA	194	8.6	69	5	10
7	225	6.9	74	5	11
16	256	9.7	69	5	12

4. Predict Ozone

Ozone	Solar.R	Wind	Temp	Month	Day
12.51	194	8.6	69	5	10
7.00	225	6.9	74	5	11
16.00	256	9.7	69	5	12

Multiple imputation in detail...

5. Set Solar.R to NA's / Regression on complete cases

Ozone	Solar.R	Wind	Temp	Month	Day
12.51	194	8.6	69	5	10
7.00	NA	6.9	74	5	11
16.00	256	9.7	69	5	12

6. Predict Solar.R

Ozone	Solar.R	Wind	Temp	Month	Day
12.51	194.00	8.6	69	5	10
7.00	201.41	6.9	74	5	11
16.00	256.00	9.7	69	5	12

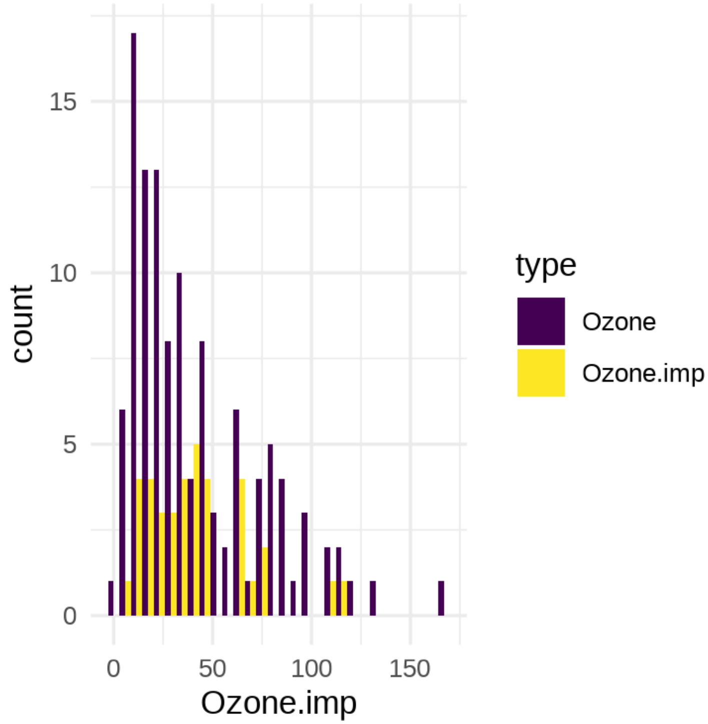
Multiple imputation in detail...

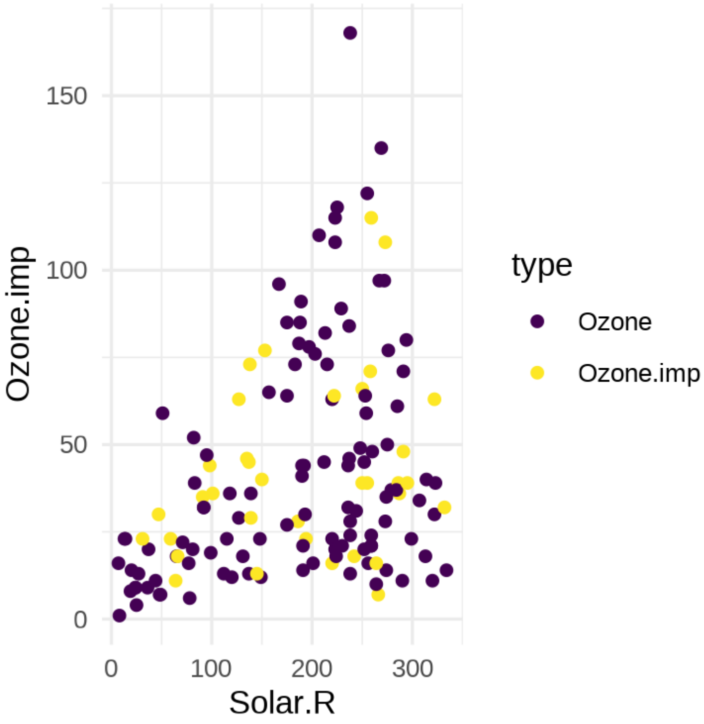
7. Set Ozone to NA's / Regression on complete cases

Ozone	Solar.R	Wind	Temp	Month	Day
NA	194.00	8.6	69	5	10
7	201.41	6.9	74	5	11
16	256.00	9.7	69	5	12

Repeat until convergence

Multiple imputation





Takeaways

- Under MCAR, one can analyze the observed observation and ignore discard any missing observations
- **Rule of thumb:** Assume MAR unless there is a good reason not to!
- Understand the missing type and data before anything (tips: missing rate, balance, correlation, data size, …)
- There is no single magical method to deal with missingness, the right choice depends on your data
- Benefit from multiple imputation to account for uncertainty
- Be vigilant in using open source packages
- Check literature for new methodologies

Thank you! Questions?

Slides: https://github.com/wittmaan/missing-data

Literature

Raghunathan, T. E., J. M. Lepkowski, J. Van Hoewyk, et al. (2001). "A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models". In: *Survey Methodology* 27, pp. 85-96.

Schafer, J. L. and J. W. Graham (2002). "Missing Data: Our View of the State of the Art". In: *Psychol Methods* 7, pp. 147-177.

vanBuuren, S. (2018). *Flexible Imputation of Missing Data*. second. Accessed: 2021-05-02. CRC Press.

vanBuuren, S. and K. Groothuis-Oudshoorn (2011). "mice: Multivariate Imputation by Chained Equations in R". In: *Journal of Statistical Software* 45, pp. 1-67.

Links

- https://www.deeplearning.ai/the-batch/issue-84/
- https://stefvanbuuren.name/publication/vanbuuren-2018/
- http://pol346.com/2021/week10_02.html#1
- https://htmlpreview.github.io/?https://raw.githubusercontent.com/ehsanx/spph504-007/master/Lab6/lab6part1.html
- https://rstudio-pubs-static.s3.amazonaws.com/445649_5f323f9cc6aa4333b404882e67e9c344.html
- https://s3.amazonaws.com/assets.datacamp.com/production/course_17404/slides/ch