Missing Data Treatments

Presentation Outline

- Types of Missing Data
- Listwise Deletion
- Pairwise Deletion
- Single Imputation Methods
  - Mean Imputation
  - Hot Deck Imputation
- Multiple Imputation
- Data Simulation
Types of Missing Data

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

**Missing Completely At Random (MCAR)**

- No relationship between the missing data and any variables
- Probability of missingness is independent of all other variables
  - Every observation is as equally likely to be missing as any another observation.
- Most missing data treatments can be performed on datasets with data MCAR without introducing bias.

**Example:**
- A student oversleeps and does not arrive in time to take the first section of a test
Missing At Random (MAR)

- No relationship between the missing data and the independent variable where the missingness occurs

- **However**, the likelihood of missingness is related to another variable in the dataset.

- Examples:
  - Women report their weight on a survey less frequently than males
  - One ethnicity reports income on a questionnaire less frequently than another ethnicity

Missing Not At Random (MNAR)

- The probability of an observation being missing depends on its measured variable.

- This is the most troublesome type of missing data and is often termed “non-ignorable.”

- Examples:
  - People who are poor are more likely not to report income on a survey.
  - Struggling readers are more likely to skip questions on a reading test.
Listwise Deletion

- Process: if any observation is missing for any participant, delete all of the data for that participant.
- Listwise deletion assumes the data are MCAR.
- Pros
  - Very easy procedure
- Cons
  - Decreases the sample size & statistical power
  - Increases standard error & widens confidence intervals

Example:

<table>
<thead>
<tr>
<th>dv</th>
<th>iv1</th>
<th>iv2</th>
<th>iv3</th>
<th>iv4</th>
</tr>
</thead>
<tbody>
<tr>
<td>80</td>
<td>50</td>
<td>NA</td>
<td>NA</td>
<td>85</td>
</tr>
<tr>
<td>95</td>
<td>45</td>
<td>53</td>
<td>100</td>
<td>75</td>
</tr>
<tr>
<td>70</td>
<td>30</td>
<td>65</td>
<td>110</td>
<td>78</td>
</tr>
<tr>
<td>NA</td>
<td>42</td>
<td>67</td>
<td>105</td>
<td>92</td>
</tr>
</tbody>
</table>
Listwise Deletion

- Example:

<table>
<thead>
<tr>
<th>dv</th>
<th>iv1</th>
<th>iv2</th>
<th>iv3</th>
<th>iv4</th>
</tr>
</thead>
<tbody>
<tr>
<td>95</td>
<td>45</td>
<td>53</td>
<td>100</td>
<td>75</td>
</tr>
<tr>
<td>70</td>
<td>30</td>
<td>65</td>
<td>110</td>
<td>78</td>
</tr>
</tbody>
</table>

Pairwise Deletion

- Process: remove cases that have missing data only when it pertains to a certain calculation.
- This is also referred to as available case analysis.
- Pairwise deletion assumes the data are MCAR.
- Pros
  - Retains more data compared with listwise deletion
- Cons
  - Can introduce bias if data are not MCAR
Pairwise Deletion

Example: If weight is not being used in the analysis, the cases where weight is missing would not be removed. If weight is a variable in the analysis, those cases would be removed.

<table>
<thead>
<tr>
<th>dv</th>
<th>age</th>
<th>weight</th>
<th>height</th>
</tr>
</thead>
<tbody>
<tr>
<td>80</td>
<td>50</td>
<td>NA</td>
<td>58</td>
</tr>
<tr>
<td>95</td>
<td>45</td>
<td>100</td>
<td>62</td>
</tr>
<tr>
<td>70</td>
<td>30</td>
<td>110</td>
<td>NA</td>
</tr>
<tr>
<td>110</td>
<td>NA</td>
<td>105</td>
<td>68</td>
</tr>
</tbody>
</table>
Single Imputation Techniques

- Imputation: substituting a value for a missing observation
- Single Imputation: each missing value is filled in with one plausible value
- Single Imputation Techniques
  - Mean Imputation
  - Hot Deck Imputation

Mean Imputation

- This technique imputes the mean of a variable for the missing observations for that variable.

**Pros**
- Retains sample size

**Cons**
- Decreases standard deviation and standard errors
- Creates smaller confidence intervals, increasing the probability of Type 1 errors
Mean Imputation

例示:

<table>
<thead>
<tr>
<th>dv</th>
<th>iv1</th>
<th>iv2</th>
<th>iv3</th>
<th>iv4</th>
</tr>
</thead>
<tbody>
<tr>
<td>80</td>
<td>50</td>
<td>NA</td>
<td>NA</td>
<td>86</td>
</tr>
<tr>
<td>95</td>
<td>45</td>
<td>54</td>
<td>100</td>
<td>76</td>
</tr>
<tr>
<td>70</td>
<td>30</td>
<td>65</td>
<td>110</td>
<td>78</td>
</tr>
<tr>
<td>NA</td>
<td>43</td>
<td>67</td>
<td>105</td>
<td>92</td>
</tr>
</tbody>
</table>

Means: 82 42 62 105 83
Hot Deck Imputation

- Process: for each missing value, find an observation with similar values in the X and take its Y value. If multiple matching values are found, the mean of those values is imputed.
- This can also be referred to as matching.
- Hot deck imputation utilizes the current dataset to find matches. Cold deck imputation utilizes an existing dataset to find matches.

Hot Deck Imputation

- Pros
  - Retains size of dataset
- Cons
  - Difficult to do when there are multiple variables with missing data
  - Reduces standard errors by underestimating the variability of the variable
**Hot Deck Imputation**

* Example:

<table>
<thead>
<tr>
<th>dv</th>
<th>iv</th>
</tr>
</thead>
<tbody>
<tr>
<td>90</td>
<td>4</td>
</tr>
<tr>
<td>NA</td>
<td>3</td>
</tr>
<tr>
<td>64</td>
<td>3.5</td>
</tr>
<tr>
<td>100</td>
<td>5</td>
</tr>
<tr>
<td>88</td>
<td>4</td>
</tr>
<tr>
<td>NA</td>
<td>6</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>dv</th>
<th>iv</th>
</tr>
</thead>
<tbody>
<tr>
<td>90</td>
<td>4</td>
</tr>
<tr>
<td>64</td>
<td>3</td>
</tr>
<tr>
<td>64</td>
<td>3.5</td>
</tr>
<tr>
<td>100</td>
<td>5</td>
</tr>
<tr>
<td>88</td>
<td>4</td>
</tr>
<tr>
<td>100</td>
<td>6</td>
</tr>
</tbody>
</table>

**Multiple Imputation**

* Process: each missing value is replaced with multiple plausible values. This creates multiple possible datasets. Then, these datasets are “pooled” together to come up with one result

**Impute**

Creates multiple possible datasets

**Analyze**

Run analysis on each dataset

**Pool**

Find average of estimates
Multiple Imputation

- Multiple methods for computing missing values
  - Predictive Mean Matching (pmm)
  - Bayesian Linear Regression (norm)
  - Logistic Regression (logreg)
  - Linear Discriminant Analysis (lda)
  - Random sample from observed values (sample)
  - Many others

Multiple Imputation

- Pros
  - Imputes multiple plausible values - reduces possibility for bias

- Cons
  - Difficult to compute
Practice in R - Setting up Data

* Create this data frame in R and name it “example”

* Run regression with Y as the DV and X as the IV

Coefficients:

| Estimate  | Std. Error | t value | Pr(>|t|) |
|-----------|------------|---------|----------|
| (Intercept) | 4.6867 | 0.9870 | 4.748 | 0.00209 ** |
| x | 0.1379 | 0.1615 | 0.854 | 0.42150 |
| --- | | | | |
| Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1 |

Residual standard error: 1.445 on 7 degrees of freedom
(3 observations deleted due to missingness)
Multiple R-squared: 0.09431,  Adjusted R-squared: -0.03508
F-statistic: 0.7289 on 1 and 7 DF,  p-value: 0.4215

Practice in R - Listwise Deletion

* Listwise Deletion

(examplelistwise<-na.omit(example))

* Run regression with y as DV and x as IV

Coefficients:

| Estimate  | Std. Error | t value | Pr(>|t|) |
|-----------|------------|---------|----------|
| (Intercept) | 4.6867 | 0.9870 | 4.748 | 0.00209 ** |
| x | 0.1379 | 0.1615 | 0.854 | 0.42150 |
| --- | | | | |
| Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1 |

Residual standard error: 1.445 on 7 degrees of freedom
Multiple R-squared: 0.09431,  Adjusted R-squared: -0.03508
F-statistic: 0.7289 on 1 and 7 DF,  p-value: 0.4215
Practice in R - Mean Imputation

✦ Mean Imputation

library(Hmisc)
exampmean<-example
exampmean$x<-impute(exampmean$x, mean)

✦ Run regression with y as DV and x as IV

Coefficients:

|             | Estimate | Std. Error | t value | Pr(>|t|) |
|-------------|----------|------------|---------|----------|
| (Intercept) | 4.4728   | 1.1004     | 4.065   | 0.00227  ** |
| x           | 0.1379   | 0.1857     | 0.743   | 0.47476  |

---

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.661 on 10 degrees of freedom
Multiple R-squared: 0.05227, Adjusted R-squared: -0.0425
F-statistic: 0.5516 on 1 and 10 DF,  p-value: 0.4748

Practice in R - Hot Deck Imputation

✦ Hot Deck Imputation

library(rrp)
examphehd<-rrp.impute(example)
examphehdd<-examphehd$new.data

✦ Run regression with y as DV and x as IV

Coefficients:

|             | Estimate | Std. Error | t value | Pr(>|t|) |
|-------------|----------|------------|---------|----------|
| (Intercept) | 4.2215   | 0.8437     | 5.003   | 0.000535  *** |
| x           | 0.2115   | 0.1528     | 1.384   | 0.196413  |

---

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.563 on 10 degrees of freedom
Multiple R-squared: 0.1608, Adjusted R-squared: -0.0425
F-statistic: 1.916 on 1 and 10 DF,  p-value: 0.1964
Practice in R - Multiple Imputation

* Multiple Imputation

```r
library(mice)
exampler <- mice(example, meth=c("","pmm"), maxit=1)
exampler2 <- with(exampler, lm(y~x))
mipooled <- pool(exampler2)
mipooled
```

* Run regression with y as DV and x as IV

|        | est     | se      | t       | df     | Pr(>|t|)   |
|--------|---------|---------|---------|--------|------------|
| (Intercept) | 5.15015978 | 1.1108854 | 4.63608574 | 7.679074 | 0.00186596 |
| x       | 0.01100627 | 0.1777149 | 0.06193217 | 7.486365 | 0.95223815 |

Practice in R - Comparing Methods

Listwise: grey
Mean Imputation: black
Hot Deck: blue
Multiple Imputation: purple
Simulation in R

- Population = 100,000
- Variables: DV, IV1, IV2, IV3
- Randomly sampled 5 subsets, n = 5,000
- Created 3 datasets from each subsets with 5%, 10%, and 20% missingness on IV1
- Performed Listwise Deletion, Mean Imputation, Hot Deck Imputation, and Multiple Imputation on each dataset
- Calculated regression estimates
- Calculated Percent Relative Parameter Bias and Relative Standard Error Bias
Comparing Methods - PRPB

- Percent Relative Parameter Bias (PRPB)
  - Measures the amount of bias introduced under a specific set of conditions (e.g., missing data treatments)
  \[
  B(\hat{\theta}_p) = \left( \frac{\hat{\theta}_p - \theta_p}{\theta_p} \right) \times 100
  \]
  \(\hat{\theta}_p\) : mean of the pth parameter for x estimates
  \(\theta_p\) : corresponding population parameter
  - Produces standardized metric to examine the size and direction of the bias
  - Values above 5% or below -5% are considered unacceptable

<table>
<thead>
<tr>
<th></th>
<th>Intercept</th>
<th>IV1</th>
<th>IV2</th>
<th>IV3</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Listwise Deletion PRPB</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5% missing</td>
<td>-1.569</td>
<td>-0.064</td>
<td>2.640</td>
<td>-4.672</td>
</tr>
<tr>
<td>10% missing</td>
<td>-1.602</td>
<td>-0.315</td>
<td>1.743</td>
<td>-2.645</td>
</tr>
<tr>
<td>20% missing</td>
<td>-1.581</td>
<td>-0.243</td>
<td>3.823</td>
<td>-3.991</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Intercept</th>
<th>IV1</th>
<th>IV2</th>
<th>IV3</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Hot Deck Imputation PRPB</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5% missing</td>
<td>-1.688</td>
<td>2.749</td>
<td>2.561</td>
<td>2.562</td>
</tr>
<tr>
<td>10% missing</td>
<td>-1.700</td>
<td>5.856</td>
<td>0.525</td>
<td>3.288</td>
</tr>
<tr>
<td>20% missing</td>
<td>-1.762</td>
<td>12.544</td>
<td>0.569</td>
<td>7.024</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Intercept</th>
<th>IV1</th>
<th>IV2</th>
<th>IV3</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Mean Imputation PRPB</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5% missing</td>
<td>-1.723</td>
<td>-0.169</td>
<td>5.743</td>
<td>4.658</td>
</tr>
<tr>
<td>10% missing</td>
<td>-1.462</td>
<td>-0.502</td>
<td>5.058</td>
<td>-11.168</td>
</tr>
<tr>
<td>20% missing</td>
<td>-0.877</td>
<td>-0.771</td>
<td>5.454</td>
<td>-46.752</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Intercept</th>
<th>IV1</th>
<th>IV2</th>
<th>IV3</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Multiple Imputation PRPB</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5% missing</td>
<td>-1.658</td>
<td>-0.281</td>
<td>3.331</td>
<td>0.692</td>
</tr>
<tr>
<td>10% missing</td>
<td>-1.544</td>
<td>-0.046</td>
<td>2.142</td>
<td>-6.233</td>
</tr>
<tr>
<td>20% missing</td>
<td>-1.519</td>
<td>-0.507</td>
<td>3.378</td>
<td>-7.736</td>
</tr>
</tbody>
</table>
Comparing Methods - PRPB

Listwise Deletion PRPB

Mean Imputation PRPB

5% missing: Grey
10% missing: Black
20% missing: Blue
Comparing Methods - PRPB

Hot Deck Imputation PRPB

Percentage Relative Parameter Bias

Parameter

Multiple Imputation PRPB

Percentage Relative Parameter Bias

Parameter

5% missing: Grey
10% missing: Black
20% missing: Blue
Comparing Methods - RSEB

• Relative Standard Error Bias (RSEB)
  • Measures the amount of bias in standard error estimates
    \[ B(s_{\hat{\theta}_p}) = \left( \frac{s_{\hat{\theta}_p} - \bar{s}_{\hat{\theta}_p}}{s_{\hat{\theta}_p}} \right) \times 100 \]
    - \( \bar{s}_{\hat{\theta}_p} \) : mean of the standard errors of the intercepts
    - \( s_{\hat{\theta}_p} \) : standard deviation of the intercepts
  • Produces standardized metric to examine the size and direction of the bias
  • Values above 10% or below -10% are considered unacceptable

Comparing Methods - RSEB

<table>
<thead>
<tr>
<th></th>
<th>Listwise</th>
<th>Mean Imputation</th>
<th>Hot Deck Imputation</th>
<th>Multiple Imputation</th>
</tr>
</thead>
<tbody>
<tr>
<td>5% missing</td>
<td>82.47</td>
<td>102.55</td>
<td>85.38</td>
<td>107.45</td>
</tr>
<tr>
<td>10% missing</td>
<td>68.77</td>
<td>86.43</td>
<td>55.62</td>
<td>39.48</td>
</tr>
<tr>
<td>20% missing</td>
<td>51.54</td>
<td>39.62</td>
<td>7.06</td>
<td>66.21</td>
</tr>
</tbody>
</table>
Comparing Methods - RSEB

RSEB for Different Data Treatment Methods

- Listwise: grey
- Mean Imputation: black
- Hot Deck: blue
- Multiple Imputation: purple

Conclusions

- Prevent missing data
- If data is missing, attempt to determine why it is missing.
- No “silver bullet” treatment method
References


