Introduction to Missing Data and Imputation

- Allison, 2001 and Horton, Kleinman, 2007

Tour the jargon of "missingness" in data in the context of a statistical analysis.

\[ \text{missing data are not the only problem, of course!} \]

Context - Statistical Learning

Data set of \( N \) observations of \( P \) variables/features \( \Rightarrow X \) (matrix)
\( N \) observations of 1 response/output \( \Rightarrow Y \) (vector)

We assume \( Y = f(X) + \epsilon \)
\( f \) unknown \( \epsilon \) mean zero, ind. of \( X \)

\[ \text{Goal: learn } f \text{ that best approximates } f \text{ in order to understand the system and/or predict new responses when given new (input) variables} \]

The classic example is linear regression: \( \hat{f}(X) = \alpha + \beta_1 X_1 + \cdots + \beta_p X_p \)

**Dataset:** Kicks Inpatient Dataset

<table>
<thead>
<tr>
<th>( X_1 ) = age</th>
<th>( X_2 ) = length of stay</th>
<th>( X_3 ) = not urgent</th>
<th>( Y ) = (non) routine discharge</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>3.5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Columns are variables
Rows are observations

Missing Data

\[ \text{One kind of error } \rightarrow \text{some variables are missing for some observations} \]
Example

\[
\begin{bmatrix}
X_1 & X_2 & X_3 \\
\text{Obs} & \text{Obs} & \text{Obs} \\
\text{Obs} & \text{Obs} & \text{M} \\
\text{Obs} & \text{M} & \text{M} \\
\text{M} & \text{obs} & \text{obs} \\
\vdots & \vdots & \vdots
\end{bmatrix}
\begin{bmatrix}
Y
\end{bmatrix}
\]

Assume all response variables are observed but 0 or more input variables may not be measured.

\[
\text{each row of } X \text{ can be partitioned into } X^{\text{obs}} \text{ and } X^{\text{obs}}
\]

\[
R = \begin{bmatrix}
1 & 1 & 1 \\
1 & 1 & 0 \\
0 & 1 & 1
\end{bmatrix}
\]

Questions we need to ask:

Why are observations missing?

\[\rightarrow\] survey respondents refuse to answer questions
\[\rightarrow\] data recording is not consistent
\[\rightarrow\] intentional exclusion for privacy

When are observations missing? Randomly?

\[\rightarrow\] Are certain combinations of observed data predictive of "missingness"?

Ex: KIDS: Longer hospital stays (Xa) correlated w/ less missing data?

\[\nabla\] Do different values of Y have more/less missing observations for Xj's?
Quantifying Randomness

Let's suppose we have a probability model to predict the values of R. \( \Pr(R|Y,X,\beta) \) creates

\[ \Rightarrow \text{Simplest model: Constant prob } \rightarrow \text{Each } R_i \text{ is } 1 \text{ w/ prob } \beta, \ 0 \text{ w/ prob. } 1-\beta \]

\[ \Rightarrow \text{Constant param.} \]

This idea (modeling R) lets us describe different scenarios for when the data may be missing. \( \Rightarrow \text{Mechanism} \)

- **Missing Completely At Random (MCAR)**

\[ \Pr(R|Y,X,\beta) = \Pr(R|Y,X^{\text{obs}},X^{\text{mix}},\beta) = \Pr(R|\beta) \]

This means that missingness is not related to any factor in the data, known or not!

- **Missing At Random (MAR)**

Here missingness may depend on observed quantities but not unobserved quantities:

\[ \Rightarrow \Pr(R|Y,X,\beta) = \Pr(R|Y,X^{\text{obs}},\beta) \]

\[ \text{x^{obs} remaining but } x^{\text{mix}} \text{ is gone!} \]

⚠️ It is impossible to tell if data are MAR! we don't know the values of the missing data so we can't compare the values of those w/ and w/o missing data to see if there is a systematic difference.

- **Non-ignorable / Missing Not At Random (MNAR)**

\[ \Pr(R|Y,X,\beta) \text{ cannot be simplified further. This is the worst case!} \]
Dealing w/ Missing Data

Can we still learn \( \hat{\beta} \) in the presence of \( X_{mis} \)?

→ What’s the simplest thing we can do?

**Listwise deletion**

also known as: case wise deletion, complete case analysis

Sounds fancy! It means throw out all observations w/ any missing values.

→ This actually works well when MCAR and research has investigated how it can work under MAR.

→ Disadvantage: are we wasting a bunch of data?

KIDS data 41% of observations have \( \geq 1 \) missing values.

**Pairwise Deletion**: (available case analysis)

Sometimes you can decompose the estimation of \( \hat{\beta} \) in such a way that you don’t look at all \( p \) variables at the same time, but instead look at \( p_{mis} \) (first we compute something on \( X_1 \) and \( X_2 \), then \( X_1 \) and \( X_3 \), then...) then pool those calculations together.

→ example: linear regression by estimating covariance matrix

When we are considering \( X_1 \) and \( X_2 \) use all observations whose both variables are present. When considering \( X_1 \) and \( X_3 \), use the (potentially) different set of observations where both variables are present.

→ Uses more of the available data, but requires MCAR and parameter estimates may be biased, b/c different sample sizes are used for different parts of the calculation.

**Dummy Variable Adjustment**

Effectively means incorporate the observed values of \( R \) into the statistical learning method.

→ biases search for \( \hat{\beta} \) to fallen out of favor!