Intro 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.

→ 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 $Y = f(X) + \epsilon$ (very general)

(f: unknown; ϵ: mean zero, indep. of X)

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

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

Dataset: Kids Inpatient Dataset

	X_1 = age	X_2 = length of stay	X_3 = urgent/not urgent	Y = (non) routine discharge
(there are more than 3 variables)	6	3.5	0	0
	9	1	1	1
	8	0	1	1
	⋮	⋮	⋮	⋮

columns are variables
rows are observations

X Y

Missing Data

→ One kind of error → some variables are missing for some observations

Example

X_1	X_2	X_3	Y
Obs	Obs	Obs	Obs
Obs	Obs	M	Obs
Obs	M	M	Obs
M	Obs	Obs	Obs
⋮	⋮	⋮	⋮

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

← each row of X can be partitioned into X^{obs} and X^{mis}

→ Can construct an indicator matrix R where $R_{ij} = 1$ if j^{th} variable (X_j) was observed for observation i:

$$R = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 1 \end{bmatrix}$$

Questions we need to ask:

Why are observations missing?

→ survey respondents refuse to answer questions
→ data recording is not consistent
→ intentional exclusion for privacy

When are observations missing? Randomly?

→ Are certain combinations (or values) of observed data predictive of "missingness"?

Ex KIDS: Longer hospital stays (X_2) correlated w/ less missing data?

⚠ Do different values of Y have more/less missing observations for X_j's.

Quantifying Randomness Let's suppose we have a probability model to predict the values of R. $\rightarrow Pr(R|Y,X,\beta)$ (β: optimal parameters)

$\Rightarrow$ Simplest model: Constant Prob $\rightarrow$ Each R_{ij} is 1 w/ prob β, 0 w/ prob. $1-\beta$

(β: parameter.)

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

- Missing Completely At Random (MCAR)

$$Pr(R|Y,X,\beta) = Pr(R|Y,X^{obs},X^{mis},\beta) = Pr(R|\beta)$$

(β: constant parameter(s))

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:

$$Pr(R|Y,X,\beta) = Pr(R|Y,X^{obs},\beta)$$

(X^{obs} remains but X^{mis} is gone!)

It is impossible to test 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)

$Pr(R|Y,X,\beta)$ cannot be simplified further. This is the worst case!

Dealing w/ Missing Data

Can we still learn $\hat{f}$ 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/ <u>any</u> missing values.

→ This actually works <u>well</u> 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 ≥ 1 missing values.

Pairwise Deletion: (available case analysis)

Sometimes you can <u>decompose</u> the estimation of $\hat{f}$ in such a way that you don't look at all p variables at the same time, but instead look at <u>pairs</u> (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 where 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 CI. on 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 $\hat{f}$ → fallen out of favor!