Missing Data

Author: Nicholas G Reich

This material is part of the **statsTeachR** *project*

Today's Lecture

- Types of missing data
- Describing your missing data
- Multiple imputation

Missing data notation

Data model

We assume we have a sample of n observations, and we are primarily interested in the conditional distribution

$$f(Y_i|\mathbf{X}_i, \beta)$$

We split $\mathbf{X}$ into two components, $\mathbf{X}^{obs}$ and $\mathbf{X}^{mis}$ for the observed and missing portions of $\mathbf{X}$, respectively.

Missing data model

We define a response indicator, $\mathbf{R}$ to denote missingness: $R_j = 1$ if X_j is observed, and 0 otherwise. Types of missingness can be categorized by how Y and $\mathbf{X}$ relate to a probability model for R:

$$p(\mathbf{R}|Y, \mathbf{X})$$

Notation adapted from Horton and Kleinman, *American Statistician*, 2007.

Types of Missing Data

Missing Completely at Random (MCAR)

- No data, observed or unobserved, are related to missingness.
- $p(\mathbf{R}|Y, \mathbf{X}) = p(\mathbf{R}|Y, \mathbf{X}^{obs}, \mathbf{X}^{mis}) = p(\mathbf{R}|\phi)$

Missing at Random (MAR)

- No unobserved data are related to missingness, but observed data could be used to predict missingness.
- $p(\mathbf{R}|Y, \mathbf{X}) = p(\mathbf{R}|Y, \mathbf{X}^{obs}, \phi)$

Missing Not at Random (MNAR) or unignorable missingness

- Missingness relationship cannot be simplified: it depends on unobserved data!
- $p(\mathbf{R}|Y, \mathbf{X}) = p(\mathbf{R}|Y, \mathbf{X})$

Testing for the different types of data

Tests about the type of data you have

- MAR vs. MNAR: Not a definitive test here. Best option is to use your domain-specific knowledge about the data.
- MCAR vs. MAR: Little's test can weigh evidence for/against these two settings.

Little's H_0: The data is MCAR

Low p-values suggest that the data are MAR; high p-values suggest they are MCAR.

```
library(openintro)
data(ncbirths)
test <- BaylorEdPsych::LittleMCAR(ncbirths)

## this could take a while

test$p.value

## [1] 0
```

Types of analyses for missing data

Analysis strategies (in rough order of desirability, low to high)

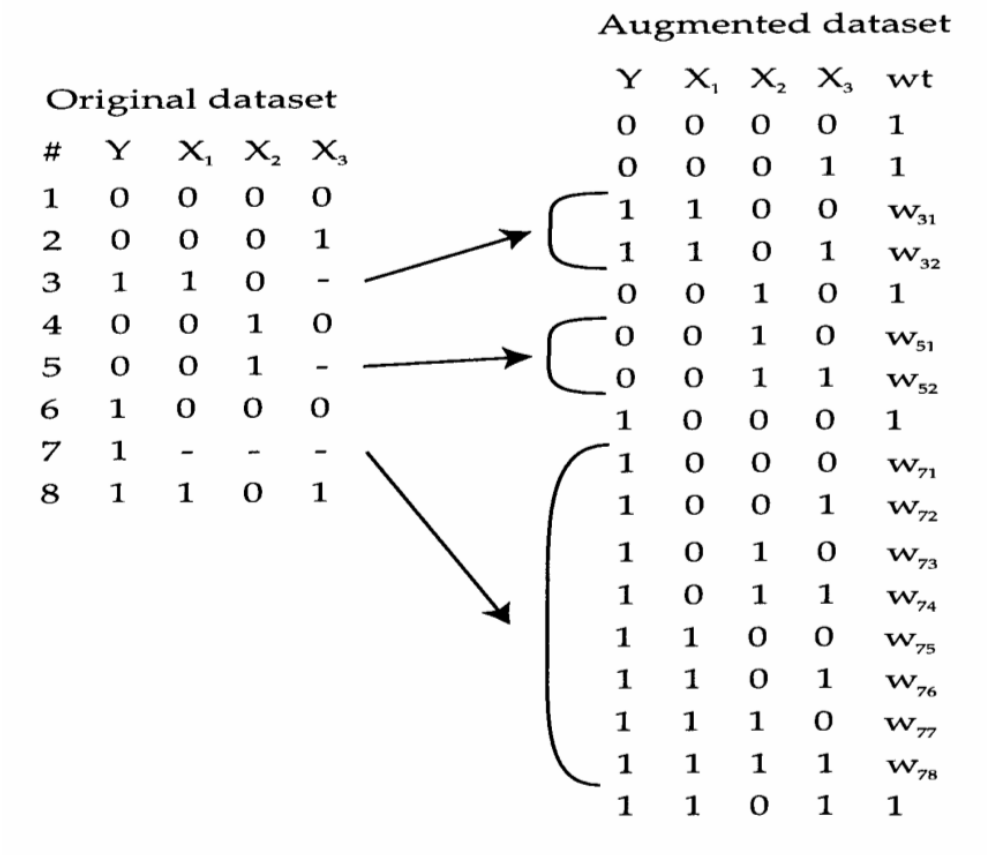
- MCAR only: Complete case a.k.a. "listwise deletion"
- Ad-hoc methods (e.g. mean imputation)
- Weighting methods
- MAR: Likelihood-based approaches (e.g. EM algorithm)
- MAR: Multiple Imputation (many flavors)
- MAR: Bayesian methods

Types of analyses for missing data

Analysis strategies (in rough order of desirability, low to high)

- MCAR only: Complete case a.k.a. "listwise deletion"
- Ad-hoc methods (e.g. mean imputation)
- Weighting methods
- MAR: Likelihood-based approaches (e.g. EM algorithm)
- MAR: Multiple Imputation (many flavors)
- MAR: Bayesian methods

Likelihood based approach



Original dataset

#	Y	X_1	X_2	X_3
1	0	0	0	0
2	0	0	0	1
3	1	1	0	-
4	0	0	1	0
5	0	0	1	-
6	1	0	0	0
7	1	-	-	-
8	1	1	0	1

Augmented dataset

Y	X_1	X_2	X_3	wt
0	0	0	0	1
0	0	0	1	1
1	1	0	0	w_{31}
1	1	0	1	w_{32}
0	0	1	0	1
0	0	1	0	w_{51}
0	0	1	1	w_{52}
1	0	0	0	1
1	0	0	0	w_{71}
1	0	0	1	w_{72}
1	0	1	0	w_{73}
1	0	1	1	w_{74}
1	1	0	0	w_{75}
1	1	0	1	w_{76}
1	1	1	0	w_{77}
1	1	1	1	w_{78}
1	1	0	1	1

Figure 2. Use of likelihood-based approach with EM algorithm to incorporate partially observed data.

EM algorithm: weights and βs are estimated iteratively.

Limitations: complicated with continuous variables

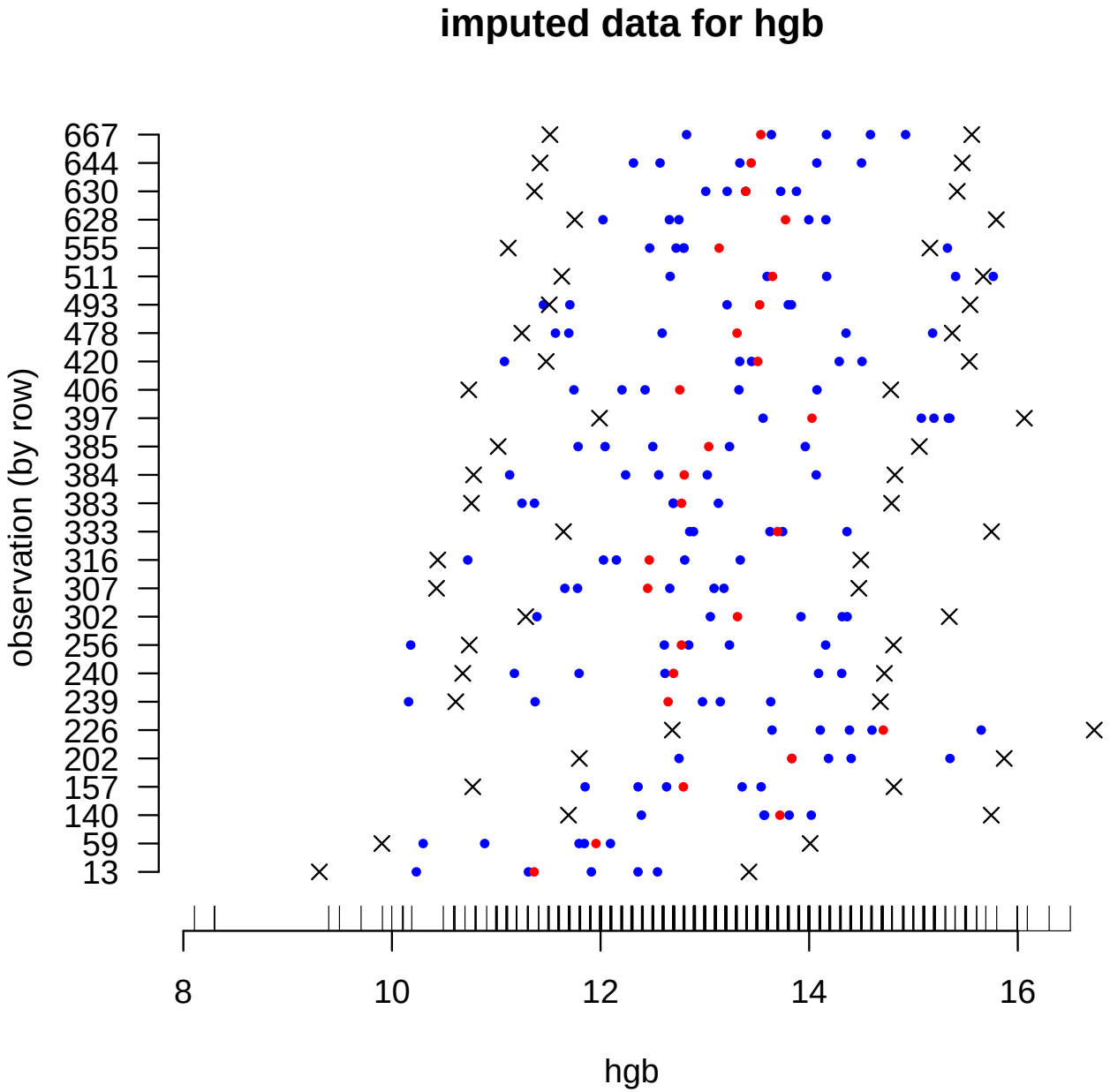
Figure credits Horton and Kleinman, *American Statistician*, 2007.

Multiple imputation

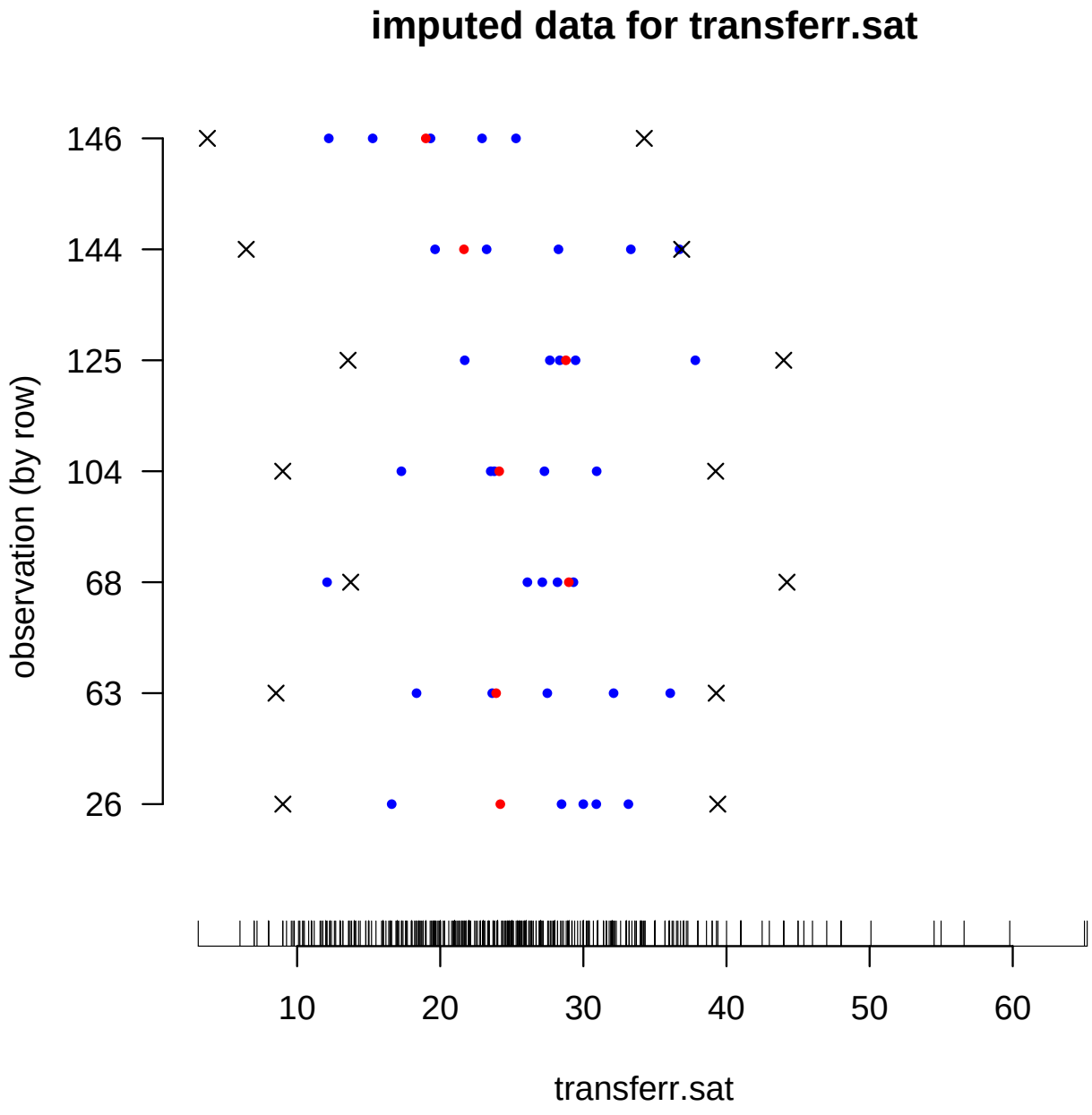
General approach

- For each missingness pattern, a model is built to use the available covariates to estimate the missing covariates.
- Random samples are taken from the predictive distribution to create multiple "complete" datasets.
- Typically, 10-15 datasets is seen as being sufficient.
- Coefficient and SE estimates are combined across datasets.

Multiple imputation: example



Multiple imputation: example



Multiple imputation results

Regression coefficients from five imputed data sets

Data set	Estimated parameter	b_0	b_1	b_2	b_3	b_4	b_5
1	Coefficient	-11.535	-2.780	1.029	-.031	-0.359	0.572
	Variance	43.204	3.323	0.013	0.013	0.013	0.012
2	Coefficient	-11.501	-4.149	1.040	-0.093	-0.583	0.876
	Variance	40.488	2.680	0.010	0.009	0.009	0.007
3	Coefficient	-10.141	-5.038	0.766	0.123	-0.252	0.625
	Variance.	42.055	3.301	0.010	0.010	0.010	0.009
4	Coefficient	-11.533	-6.920	0.870	0.084	-0.458	0.815
	Variance	28.751	1.796	0.081	0.007	0.007	0.007
5	Coefficient	-14.586	-1.115	0.718	0.050	-0.373	0.814
	Variance	32.856	2.362	0.009	0.009	0.009	0.008
	Mean b_i	-11.859	-4.000	0.885	0.027	-0.405	0.740
	Mean Var.$(\bar{W})$	37.471	2.692	0.025	0.010	0.010	0.009
	Var. of b_i (B)	2.682	4.859	0.022	0.008	0.015	0.018
	T						
	$\sqrt{T}$	40.69	8.523	0.051	0.020	0.028	0.031
		6.379	2.919	0.226	0.141	0.167	0.176
	t	-1.859	-1.370	3.916*	0.191	2.425*	4.204*

* $p < .05$ "Var." refers to the squared standard error of the coefficient.

DC Howell, Treatment of Missing Data – Part II.

Multiple imputation results

Regression coefficients from five imputed data sets

Data set	Estimated parameter	b_0	b_1	b_2	b_3	b_4	b_5
1	Coefficient	-11.535	-2.780	1.029	-.031	-0.359	0.572
	Variance	43.204	3.323	0.013	0.013	0.013	0.012
2	Coefficient	-11.501	-4.149	1.040	-0.093	-0.583	0.876
	Variance	40.488	2.680	0.010	0.009	0.009	0.007
3	Coefficient	-10.141	-5.038	0.766	0.123	-0.252	0.625
	Variance.	42.055	3.301	0.010	0.010	0.010	0.009
4	Coefficient	-11.533	-6.920	0.870	0.084	-0.458	0.815
	Variance	28.751	1.796	0.081	0.007	0.007	0.007
5	Coefficient	-14.586	-1.115	0.718	0.050	-0.373	0.814
	Variance	32.856	2.362	0.009	0.009	0.009	0.008
	Mean b_i	-11.859	-4.000	0.885	0.027	-0.405	0.740
	Mean Var.$(\bar{W})$	37.471	2.692	0.025	0.010	0.010	0.009
	Var. of b_i (B)	2.682	4.859	0.022	0.008	0.015	0.018
	T						
	$\sqrt{T}$	40.69	8.523	0.051	0.020	0.028	0.031
		6.379	2.919	0.226	0.141	0.167	0.176
	t	-1.859	-1.370	3.916*	0.191	2.425*	4.204*

* $p < .05$ "Var." refers to the squared standard error of the coefficient.

DC Howell, Treatment of Missing Data – Part II.

Best practices

Hard to argue with an approach that does the following:

- quantify the completeness of covariate data
- provide details about your approach for handling missing data
- present and discuss patterns of or reasons for missing data

Proposed guidelines for reporting missing covariate data (Burton and Altman 2004)

Summary

You will have practice with missing data methods, most importantly multiple imputation, in Lab 5!