Praktikum z ekonometrie

VŠE Praha

Tomáš Formánek

Outline

The nature of missing data

Missing completely at random ($MCAR$)

- The probability that an observation X_i is missing is unrelated to the value of X_i or to the value of any other variables.
- Any piece of data is equally likely to be missing.
- Analyses based on data with $MCAR$ observations remain unbiased. We may lose power (increased standard errors), but the estimated parameters are not biased by the absence of data.

Missing at random (MAR)

- Data meets the requirement that missingness does not depend on the value of X_i after controlling for another variable in our analysis.
- For example, data are MCAR in a specific (demographic) subgroup.

Missing Not at Random ($MNAR$)

- Missigness of X_i depends on its value (e.g. income in surveys)
- The only way to obtain an unbiased estimates of (regression) parameters is to model the missingness.

Traditional treatment of missing data

Listwise deletion (complete cases analysis)

- We omit all rows with missing data – missing information for at least one variable in the i-th individual observation. Then, we run our analyses on the observations that remain. This often results in a substantial decrease in sample size. Under the assumption that data are missing completely at random, LRMs lead to unbiased parameter estimates – still, we lose power due to exclusion of (potentially large number of) observations.

R code

```
newData <- data[complete.cases(data)==T, ]
# data is a data.frame
# or
newData <- na.omit(data)
```

Traditional treatment of missing data

Hot deck imputation

- Historically used by the US Census Bureau (since 1950's). Respondent's missing data were replaced by observed replacement data – drawn at random from a group of similar participants. Suitable, given only a few missing observations need to be replaced and given the draw is random.

Traditional treatment of missing data

Mean substitution

✓ Simple

✗ In simple linear regression models (SLRMs), this adds no new information but increases sample size – that leads to underestimated standard errors only.

Example: Data on salary and citation level of publications. 62 cases with complete data and 7 cases for which the citation index was missing. Correlations and regression coefficients were compared as follows:

Analysis	n	$corr$	$\widehat{\beta}_1$	$s.e.(\widehat{\beta}_1)$
Complete cases only	62	.55	310.747	60.95
With mean substitution	69	.54	310.747	59.12

Traditional treatment of missing data

Regression substitution

- Uses linear regression (auxiliary LRM) to predict what the missing values of regressors should be, on the basis of other variables that are present.
- For SLRMs, the same problem of error variance as in mean substitution remains. We do not add more information but we increase the sample size and (spuriously) reduce the standard error.
- May be useful for MLRMs.

Stochastic regression substitution

- This approach adds a randomly sampled residual term from the normal (or other) distribution to each value estimated by regression substitution. Adding a bit of random error to each substitution reduces, but does not eliminate, the problem of spurious reduction of the standard errors.

Maximum Likelihood Expectation-Maximization

- Computationally complex, maximum likelihood approach to the estimation of missing values Many approaches exist (e.g. the Expectation-Maximization algorithm)

  `https://www.uvm.edu/~dhowell/StatPages/Missing_Data/Missing-Part-Two.html`

Multiple Imputation (MI)

R: {mice} , {mi} , {Amelia} , …

MI motivation and algorithm

- Create several (say, 5) "imputed" values for each missing value X_{ij}. Each of the (5) versions of imputed data values are estimated/predicted using a separate ML approach from the data frame observed (different "model" is used for each imputation).
- For SLRMs, this may be simplified into mean substitution augmented by adding random errors which reflect sampling variability of X_j.

Multiple Imputation (contnd.)

How do we analyze estimates on data with MI?

1. We use each set of imputed values to form a separate completed dataset (e.g., we get 5 datasets).
2. For each completed dataset, a standard analysis (LRM) can be run.
3. Inferences can be combined across MI-based datasets.

Modern Approaches to missing data

Multiple Imputation (contnd.)

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	40.69	8.523	0.051	0.020	0.028	0.031
	$\sqrt{T}$	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.

`https://www.uvm.edu/~dhowell/StatPages/Missing_Data/Missing-Part-Two.html`

Missing dependent variable data

Special considerations apply to missing dependent variable data

- If we can assume that data are missing completely at random *(MCAR)*, we will lose power because of smaller sample sizes, but we will not have problems with biased estimates.
- If data are missing not at random *(MNAR)*, the **only way to obtain an unbiased estimate of parameters is to model missingness**. In other words we need to use a model that accounts for the missing data.
- Broadly speaking, such models are:
    - Censored Regression Models (e.g. duration analysis)
    - Truncated Regression Models
    - Sample Selection Correction models (Heckit)
    - ...