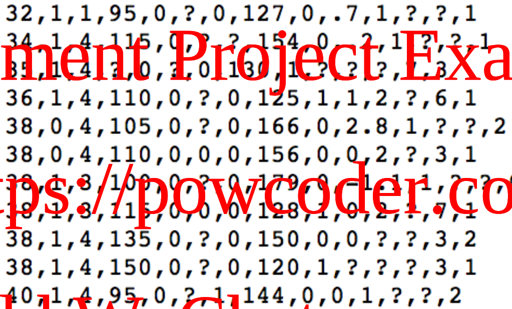
Data Cleansing — 2

Faculty of Information Technology, Monash University, Australia

FIT5196 week 7

Missing Value



```
32,1,1,95,0,?,0,127,0,.7,1,?,?,1
34,1,4,115,0,?,?,154,0,.2,1,?,?,1
35,1,4,?,0,?,0,130,1,?,?,?,?,3
36,1,4,110,0,?,0,125,1,1,2,?,6,1
38,0,4,105,0,?,0,166,0,2.8,1,?,?,2
38,0,4,110,0,0,0,156,0,0,2,?,3,1
38,1,3,100,0,?,0,179,0,-1.1,1,?,?,0
38,1,3,115,0,0,0,128,1,0,2,?,7,1
38,1,4,135,0,?,0,150,0,0,?,?,3,2
38,1,4,150,0,?,0,120,1,?,?,?,3,1
40,1,4,95,0,?,1,144,0,0,1,?,?,2
```

Missing values in the Switzerland heart disease data set are indicated by "?".

Reasons for Missing Values

- Equipment errors
- Absence of survey participants
- Unavailability in GPS signals in rural area.
- Change of circumstances: Such as death, graduation, etc.
- Filter question when a set of questions in a survey that is only asked to participants who indicate they are married.

Consequences of Missing Values

- Why is missing data a problem in data analysis?
  - All standard statistical methods presume complete information for all the variables included in analysis.

- Consequences: Ignoring or inappropriately handling missing data may lead to
  - biased estimation: over/under estimated sample mean and variance
  - Incorrect inferences/results: garbage in garbage out

- “The only really good solution to the missing data problem is not to have any. So in the design and execution of research projects, it is essential to put great effort into minimising the occurrence of missing data. Statistical adjustments can never make up for sloppy research” — Paul D. Allison

Outline

1. Missing Data Mechanisms

2. Missing Data Pattern

3. Methods for Handling Missing Values

4. Summary

Missing Data Mechanisms

- Describe relationships between measured variables and the probability of missing data
- Deciding upon the method for analysing missing values requires understanding about both the reasons for the missing values and the nature of the data for the missing observations.
- Three different missingness mechanisms:
  - Missing at random
  - Missing completely at random
  - Missing not at random

Mechanisms: Missing at Random (MAR)

- MAR: the probability of missing data on a variable is related to some other measured variable (or variables) in the analysis model but not to the values of the variable itself.
    - B: a binary $n \times p$ matrix indicating the missingness of the data
    - $Y = (Y_{obs}, Y_{miss})$
        - Y_{obs}: observed part of Y
        - Y_{miss}: missing part of Y
    - η: some unknown parameter

$$p(B \mid Y_{obs}, Y_{miss}, \eta) = p(B \mid Y_{obs}, \eta)$$

which says the probability of missingness depends on the observed portion of data via some parameter η that relates Y_{obs} to R.

- Practical issue: no way to confirm that the probability of missing data on Y is solely a function of other measured variables.

Mechanisms: Missing at Random (MAR)

- Examples
  - A psychologist is studying quality of life in a group of cancer patients and finds that elderly patients and patients with less education have a higher propensity to refuse the quality of life questionnaire.
    - The missingness in the quality of life is related to the age and education
  - An educational researcher is studying reading achievement and finds that Hispanic students have a higher rate of missing data than Caucasian students
    - The missingness in reading achievement is related to the ethic groups of students.

Mechanisms: Missing Completely at Random (MCAR)

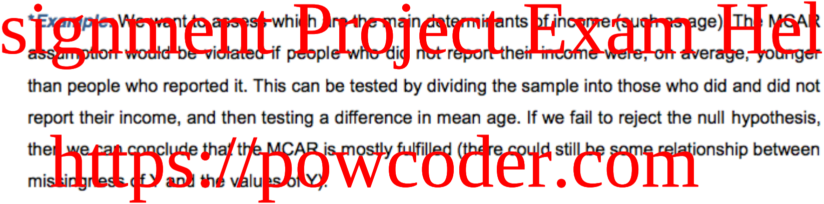
- MCAR: the probability of missing data on a variable is unrelated to other measured variables and is unrelated to the values of the variable itself.
  - B: a binary $n \times p$ matrix indicating the missingness of the data
  - $Y = (Y_{obs}, Y_{miss})$
    - Y_{obs}: observed part of Y
    - Y_{mis}: missing part of Y
  - η: some unknown parameter
  - MCAR is defined probabilistically as

    $$p(B \mid Y_{obs}, Y_{miss}, \eta) = p(B \mid \eta)$$

    which says that some parameter still governs the probability that B takes on a value of zero or one, but missingness is no longer related to the data.
- MCAR is a more restrictive condition than MAR.
- Both MAR and MCAR could be ignorable.

Mechanisms: Missing Completely at Random (MCAR)

- Example:



Example: We want to assess which are the main determinants of income (such as age). The MCAR assumption would be violated if people who did not report their income were, on average, younger than people who reported it. This can be tested by dividing the sample into those who did and did not report their income, and then testing a difference in mean age. If we fail to reject the null hypothesis, then we can conclude that the MCAR is mostly fulfilled (there could still be some relationship between missingness of Y and the values of Y).

Example adopted from "Dealing with missing data: Key assumptions and methods for applied analysis" by Marina Soley-Bori.

Mechanisms: Missing Completely at Random (MCAR)

- Effect of MCAR:

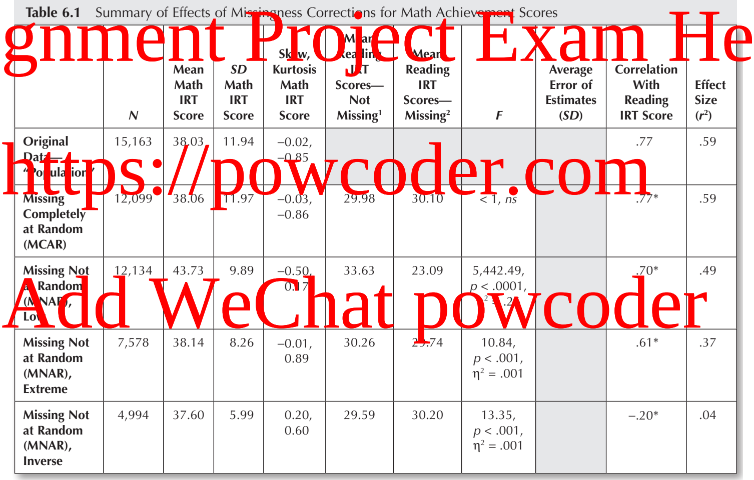


Table 6.1 Summary of Effects of Missingness Corrections for Math Achievement Scores

	N	Mean Math IRT Score	*SD* Math IRT Score	Skew, Kurtosis Math IRT Score	Mean Reading IRT Scores—Not Missing[1]	Mean Reading IRT Scores—Missing[2]	*F*	Average Error of Estimates (*SD*)	Correlation With Reading IRT Score	Effect Size (r^2)
Original Data—"Population"	15,163	38.03	11.94	−0.02, −0.85					.77	.59
Missing Completely at Random (MCAR)	12,099	38.06	11.97	−0.03, −0.86	29.98	30.10	< 1, *ns*		.77*	.59
Missing Not at Random (MNAR), Low	12,134	43.73	9.89	−0.50, 0.17	33.63	23.09	5,442.49, $p < .0001$, η^2 = .2[illegible]		.70*	.49
Missing Not at Random (MNAR), Extreme	7,578	38.14	8.26	−0.01, 0.89	30.26	29.74	10.84, $p < .001$, $\eta^2 = .001$		.61*	.37
Missing Not at Random (MNAR), Inverse	4,994	37.60	5.99	0.20, 0.60	29.59	30.20	13.35, $p < .001$, $\eta^2 = .001$		−.20*	.04

Figure is from "Dealing with missing or incomplete data" .

Mechanisms: Missing Completely at Random (MCAR)

- Test MCAR: separate the missing and the complete cases on a particular variable and examine group mean differences on other variables in the data set.
  - Univariate T-test Comparisons: It separates the missing and the complete cases on a particular variable and uses a T-test to examine group mean differences on other variables in the data set.
    - A non-significant t test: the data are MCAR.
    - A significant T statistic (or alternatively, a large mean difference): the data are MAR or MNAR.
  - Little's MCAR Test: A multivariate extension of the t-test approach that simultaneously evaluates mean differences on every variable in the data set
    - A global test of MCAR that applies to the entire data set

Mechanisms: Missing Not at Random (MNAR)

- MNAR: the probability of missing data on a variable is related to the values of the variable itself, even after controlling for other variables
  - B: a binary $n \times p$ matrix indicating the missingness of the data
  - $Y = (Y_{obs}, Y_{miss})$
    - Y_{obs}: observed part of Y
    - Y_{mis}: missing part of Y
  - η: some unknown parameter
  - MCAR is defined probabilistically as

$$p(B \mid Y_{obs}, Y_{miss}, \eta)$$

Mechanisms: Missing Not at Random (MNAR)

- Examples
  - Students with poor reading skills have missing test scores because they experienced reading comprehension difficulties during the exam.
    - The missingness in reading achievement is related to reading skills.
  - A number of patients in the cancer trial become so ill (e.g., their quality of life becomes so poor) that they can no longer participate in the study.
    - The missingness in the quality of life is related to the quality of life itself.

Mechanisms: Missing Not at Random (MNAR)

- Effects of MNAR

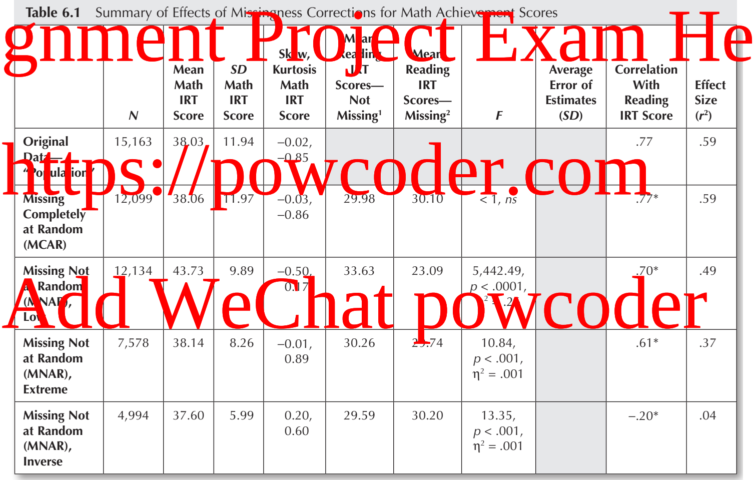
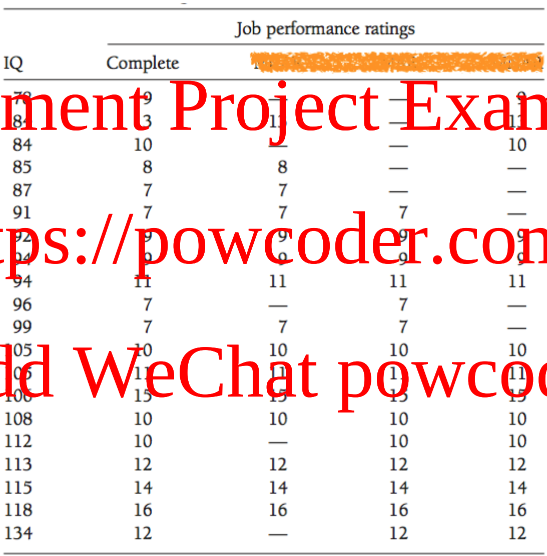


Table 6.1 Summary of Effects of Missingness Corrections for Math Achievement Scores

	N	Mean Math IRT Score	*SD* Math IRT Score	Skew, Kurtosis Math IRT Score	Mean Reading IRT Scores—Not Missing[1]	Mean Reading IRT Scores—Missing[2]	*F*	Average Error of Estimates (*SD*)	Correlation With Reading IRT Score	Effect Size (r^2)
Original Data—"Population"	15,163	38.03	11.94	–0.02, –0.85					.77	.59
Missing Completely at Random (MCAR)	12,099	38.06	11.97	–0.03, –0.86	29.98	30.10	< 1, *ns*		.77*	.59
Missing Not at Random (MNAR), Low	12,134	43.73	9.89	–0.50, 0.17	33.63	23.09	5,442.49, $p < .0001$, $\eta^2 = .2$[illegible]		.70*	.49
Missing Not at Random (MNAR), Extreme	7,578	38.14	8.26	–0.01, 0.89	30.26	29.74	10.84, $p < .001$, $\eta^2 = .001$		.61*	.37
Missing Not at Random (MNAR), Inverse	4,994	37.60	5.99	0.20, 0.60	29.59	30.20	13.35, $p < .001$, $\eta^2 = .001$		–.20*	.04

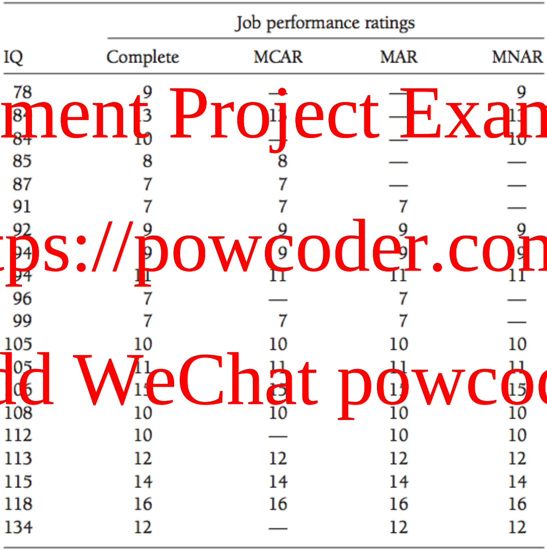
Figure is from "Dealing with missing or incomplete data" .

MAR, MCAR v.x. MNAR?



	Job performance ratings			
IQ	Complete	[redacted]	[redacted]	[redacted]
78	9	—	—	9
84	13	13	—	13
84	10	—	—	10
85	8	8	—	—
87	7	7	—	—
91	7	7	7	—
92	9	9	9	9
94	9	9	9	9
94	11	11	11	11
96	7	—	7	—
99	7	7	7	—
105	10	10	10	10
105	11	11	11	11
106	15	15	15	15
108	10	10	10	10
112	10	—	10	10
113	12	12	12	12
115	14	14	14	14
118	16	16	16	16
134	12	—	12	12

MAR, MCAR v.x. MNAR?



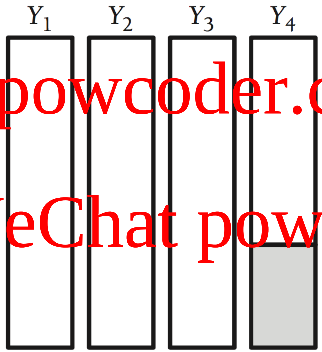
	Job performance ratings			
IQ	Complete	MCAR	MAR	MNAR
78	9	—	—	9
84	13	13	—	13
84	10	—	—	10
85	8	8	—	—
87	7	7	—	—
91	7	7	7	—
92	9	9	9	9
94	9	9	9	9
94	11	11	11	11
96	7	—	7	—
99	7	7	7	—
105	10	10	10	10
105	11	11	11	11
106	15	15	15	15
108	10	10	10	10
112	10	—	10	10
113	12	12	12	12
115	14	14	14	14
118	16	16	16	16
134	12	—	12	12

Example adopted from "Applied Missing Data Analysis" by Craig K. Enders.

Missing data Patten

A **missing data pattern** refers to the configuration of observed and missing values in a data set.

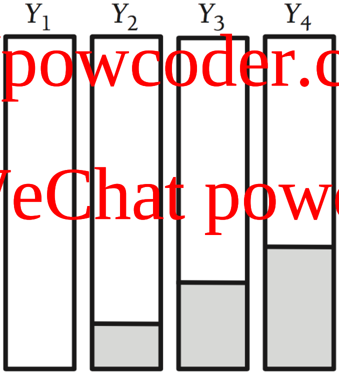
- The **univariate pattern** has missing values isolated to a single variable.



Missing data Patten

A **missing data pattern** refers to the configuration of observed and missing values in a data set.

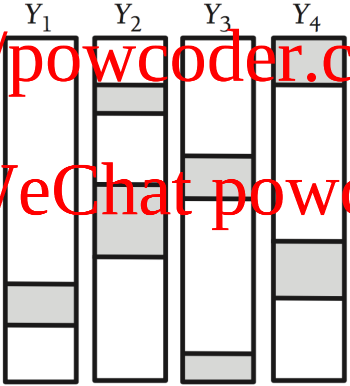
- A **monotone missing data** pattern is typically associated with a longitudinal study where participants drop out and never return.



Missing data Patten

A **missing data pattern** refers to the configuration of observed and missing values in a data set.

- a **general pattern** has missing values dispersed throughout the data matrix in a haphazard fashion.



Missing data Patten

A **missing data pattern** refers to the configuration of observed and missing values in a data set.

- Example: A study examining the effects of a program to increase students' knowledge of their asthma. It is interested in examining how a measure of a student's self efficacy beliefs about controlling their asthma symptoms relates to a number of predictors.

Variable	Definition	Possible values	*M*	*(SD)*	*N*
Asthma belief Survey	Level of confidence in controlling asthma	Range from 1, little confidence to 5, lots of confidence	4.057	(0.713)	154
Group	Treatment or control group	0 = Treatment 1 = Control	0.558	(0.498)	154
Symsev	Severity of asthma symptoms in 2-week period post-treatment	0 = no symptoms 1 = mild symptoms 2 = moderate symptoms 3 = severe symptoms	0.23[illegible]	(0.370)	141
Reading	Standardized state reading test score	Grade equivalent scores, ranging from 1.10 to 8.10	3.443	(1.636)	79
Age	Age of child in years	Range from 8 to 14	10.586	(1.605)	152
Gender	Gender of child	0 = Male 1 = Female	0.442	(0.498)	154
Allergy	Number of allergies reported	Range from 0 to 7	2.783	(1.919)	83

Symsev	Reading	Age	Allergy	# of cases	% of cases
O	O	O	O	19	12.3
M	O	O	O	1	0.6
O	M	O	O	54	35.1
O	O	O	M	56	36.4
M	M	O	O	9	5.8
M	O	O	M	1	0.6
O	M	O	M	10	6.5
O	O	M	M	2	1.3
M	M	O	M	2	1.3
# missing 13 (8.4%)	# missing 75 (48.7%)	# missing 2 (1.3%)	# missing 71 (46.1)	154	

Figures are from "A review of methods for missing data" by Pigott

Outline

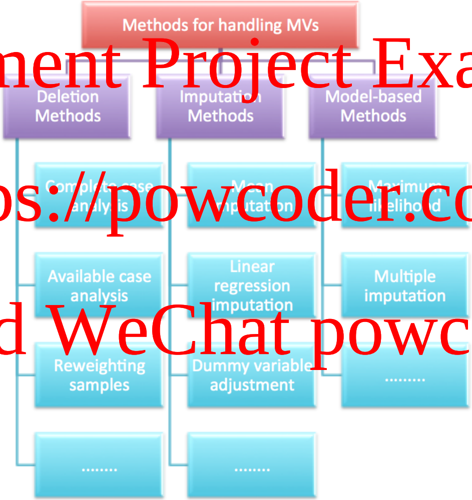
1 Missing Data Mechanisms

2 Missing Data Pattern

3 Methods for Handling Missing Values

4 Summary

Methods for handling missing values

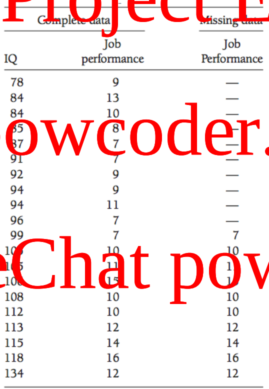


Deletion method: List-wise Deletion

- **Listwise deletion** (also known as **complete-case analysis**) discards the data for any case that has one or more missing values.

Deletion method: List-wise Deletion

- **Listwise deletion** (also known as **complete-case analysis**) discards the data for any case that has one or more missing values.



Complete data		Missing data
IQ	Job performance	Job Performance
78	9	—
84	13	—
84	10	—
85	8	—
87	7	—
91	7	—
92	9	—
94	9	—
94	11	—
96	7	—
99	7	7
105	10	10
105	11	11
106	15	15
108	10	10
112	10	10
113	12	12
115	14	14
118	16	16
134	12	12

Figure is from "Applied Missing Data Analysis"

Deletion method: List-wise Deletion

- **Listwise deletion** (also known as **complete-case analysis**) discards the data for any case that has one or more missing values.

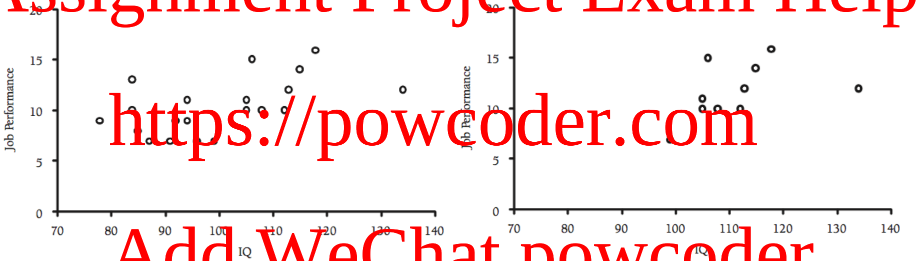


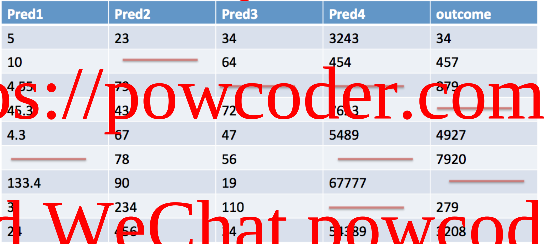
Figure is from "Applied Missing Data Analysis"

Deletion method: List-wise Deletion

- **Listwise deletion** (also known as **complete-case analysis**) discards the data for any case that has one or more missing values.
- Considerations
  - The primary benefit of list-wise deletion is convenience, producing a common set of cases for all analyses.
  - It assumes MCAR data and can produce distorted parameter estimates when this assumption does not hold.
  - Deleting the incomplete data records can produce a dramatic reduction in the total sample size, the magnitude of which increases as the missing data rate or number of variables increases.

Deletion method: Pairwise Deletion

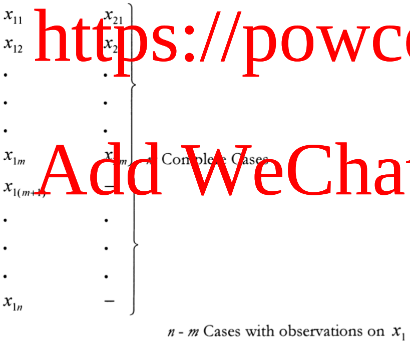
- **Pairwise deletion** (also known as **available-case analysis**) attempts to mitigate the loss of data by eliminating cases on an analysis-by-analysis basis.



Pred1	Pred2	Pred3	Pred4	outcome
5	23	34	3243	34
10	—	64	454	457
4.55	73	—	—	879
45.3	43	72	7623	—
4.3	67	47	5489	4927
—	78	56	—	7920
133.4	90	19	67777	—
3	234	110	—	279
24	456	34	54389	3208

Deletion method: Pairwise Deletion

- **Pairwise deletion** (also known as **available-case analysis**) attempts to mitigate the loss of data by eliminating cases on an analysis-by-analysis basis.
- Example: compute covariance



$$\bar{x}_1 = \sum_{i=1}^{n} x_{1i}$$

$$\bar{x}_2 = \sum_{i=1}^{m} x_{2i}$$

$$s_1^2 = \frac{\sum_{i=1}^{n} (x_{1i} - \bar{x}_1)^2}{n-1}$$

$$s_2^2 = \frac{\sum_{i=1}^{m} (x_{2i} - \bar{x}_2)^2}{m-1}$$

$$r_{xy}^2 = \frac{1}{m-1} \frac{\sum_{i=1}^{m} (x_{1i} - \bar{x}_{1(m)})(x_{2i} - \bar{x}_2)}{s_{1(m)} \, s_2}$$

Figure are from "A Review of Methods for Missing Data" by Therese D. Pigott

Deletion method: Pairwise Deletion

- **Pairwise deletion** (also known as **available-case analysis**) attempts to mitigate the loss of data by eliminating cases on an analysis-by-analysis basis.
- Considerations:
  - It requires MCAR data and can produce distorted parameter estimates when this assumption does not hold.
  - It is dependent on the magnitude of correlations that exist between variables.
  - It can produce estimated covariance matrices are outside of the range of 1.0 to 1.0, which causes estimation problems for multivariate analyses that use a covariance matrix as input data.
  - It is lack of a consistent sample base: cause problems in computing standard errors and covariance.

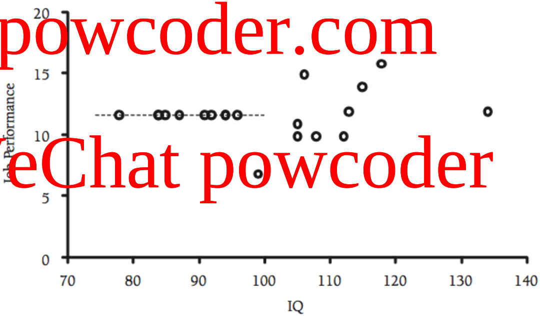
Single Imputation methods

- Single imputation: generates a single replacement value for each missing data point.
  - Yields a complete data set
  - Produces biased parameter estimates
  - Underestimates standard errors
- Methods
  - Mean Imputation
  - Regression Imputation
  - Stochastic Regression Imputation

Arithmetic mean imputation

Arithmetic mean imputation (also referred to as **mean substitution**) takes the seemingly appealing tack of filling in the missing values with the arithmetic mean of the available cases.

	Complete data	Missing data
IQ	Job performance	Job Performance
78	9	—
84	13	—
84	10	—
85	8	—
87	7	—
91	7	—
92	9	—
94	9	—
94	11	—
96	7	—
99	7	7
105	10	10
105	11	11
106	15	15
108	10	10
112	10	10
113	12	12
115	14	14
118	16	16
134	12	12



$$\mu_{complete} = 10.35,\ \mu_{miss} = 11.7,\ \mu_{impute} = 11.7$$

Regression imputation

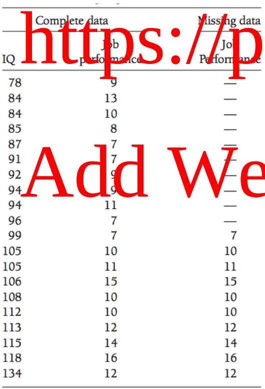
Regression imputation replaces missing values with predicted scores from a regression equation.

- Basic idea: use information from the complete variables to fill in the incomplete variables.
- Two steps:
  1. Estimate a set of regression equations that predict the incomplete variables from the complete variables.
  2. Generate predicted values for the incomplete variables

Regression imputation

Regression imputation replaces missing values with predicted scores from a regression equation.

- Basic idea: use information from the complete variables to fill in the incomplete variables.
- Example



Complete data		Missing data
IQ	Job performance	Job Performance
78	9	—
84	13	—
84	10	—
85	8	—
87	7	—
91	7	—
92	9	—
94	9	—
94	11	—
96	7	—
99	7	7
105	10	10
105	11	11
106	15	15
108	10	10
112	10	10
113	12	12
115	14	14
118	16	16
134	12	12

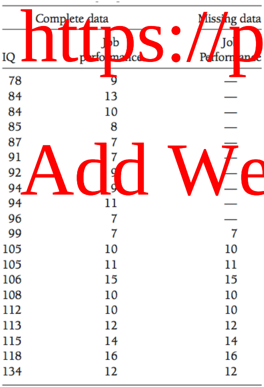
- Regression function

$$\begin{aligned} JP_i &= \beta_0 + \beta_1(IQ_i) \\ &= -2.065 + 0.123(IQ_i) \end{aligned}$$

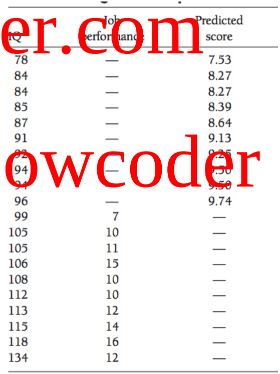
Regression imputation

Regression imputation replaces missing values with predicted scores from a regression equation.

- Basic idea: use information from the complete variables to fill in the incomplete variables.
- Example



	Complete data	Missing data
IQ	Job performance	Job Performance
78	9	—
84	13	—
84	10	—
85	8	—
87	7	—
91	7	—
92	9	—
94	9	—
94	11	—
96	7	—
99	7	7
105	10	10
105	11	11
106	15	15
108	10	10
112	10	10
113	12	12
115	14	14
118	16	16
134	12	12



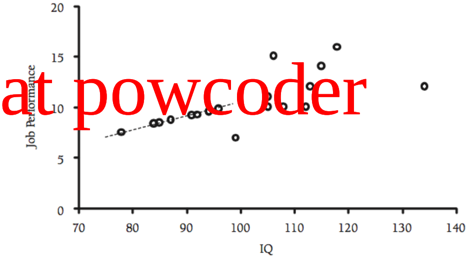
IQ	Job performance	Predicted score
78	—	7.53
84	—	8.27
84	—	8.27
85	—	8.39
87	—	8.64
91	—	9.13
92	—	9.25
94	—	9.50
94	—	9.50
96	—	9.74
99	7	—
105	10	—
105	11	—
106	15	—
108	10	—
112	10	—
113	12	—
115	14	—
118	16	—
134	12	—

Regression imputation

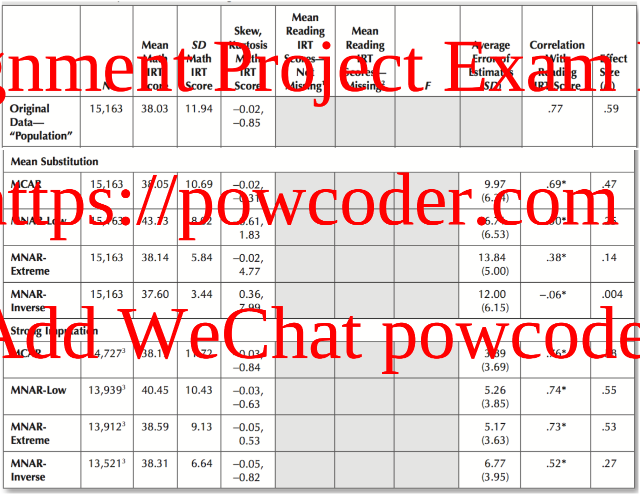
Regression imputation replaces missing values with predicted scores from a regression equation.

- Basic idea: use information from the complete variables to fill in the incomplete variables.
- Example

IQ	Job performance	Predicted score
78	—	7.53
84	—	8.27
84	—	8.27
85	—	8.39
87	—	8.64
91	—	9.13
92	—	9.25
94	—	9.50
94	—	9.50
96	—	9.74
99	7	—
105	10	—
105	11	—
106	15	—
108	10	—
112	10	—
113	12	—
115	14	—
118	16	—
134	12	—



Effects of mean and regressing imputation



	N	Mean Math IRT Score	*SD* Math IRT Score	Skew, Kurtosis Math IRT Score	Mean Reading IRT Scores—Not Missing[1]	Mean Reading IRT Scores—Missing[2]	*F*	Average Error of Estimates (*SD*)	Correlation With Reading IRT Score	Effect Size (r^2)
Original Data—“Population”	15,163	38.03	11.94	−0.02, −0.85					.77	.59
Mean Substitution										
MCAR	15,163	38.05	10.69	−0.02, −0.31				9.97 (6.34)	.69*	.47
MNAR-Low	15,163	43.73	8.02	−0.61, 1.83				6.71 (6.53)	.50*	.25
MNAR-Extreme	15,163	38.14	5.84	−0.02, 4.77				13.84 (5.00)	.38*	.14
MNAR-Inverse	15,163	37.60	3.44	0.36, 7.99				12.00 (6.15)	−.06*	.004
Strong Imputation										
MCAR	14,727[3]	38.15	11.72	−0.03, −0.84				3.89 (3.69)	.76*	.58
MNAR-Low	13,939[3]	40.45	10.43	−0.03, −0.63				5.26 (3.85)	.74*	.55
MNAR-Extreme	13,912[3]	38.59	9.13	−0.05, 0.53				5.17 (3.63)	.73*	.53
MNAR-Inverse	13,521[3]	38.31	6.64	−0.05, −0.82				6.77 (3.95)	.52*	.27

Figures are from "Dealing with missing or incomplete data"

Stochastic regression imputation

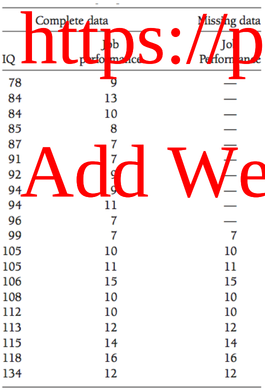
Stochastic regression imputation add random residuals to the predicate values generated by standard regression imputation.

- Basic idea: to restore lost variability to the data and effectively eliminate the biases associated with standard regression imputation methods.
- Two steps:
  1. Estimate a set of regression equations that predict the incomplete variables from the complete variables.
  2. Generate predicted values for the incomplete variables
  3. Add a normally distributed residual term to each predicted score

Stochastic regression imputation

Stochastic regression imputation add random residuals to the predicate values generated by standard regression imputation.

- Basic idea: to restore lost variability to the data and effectively eliminate the biases associated with standard regression imputation methods.
- Example:



	Complete data	Missing data
IQ	Job performance	Job Performance
78	9	—
84	13	—
84	10	—
85	8	—
87	7	—
91	7	—
92	9	—
94	9	—
94	11	—
96	7	—
99	7	7
105	10	10
105	11	11
106	15	15
108	10	10
112	10	10
113	12	12
115	14	14
118	16	16
134	12	12

- Regression function

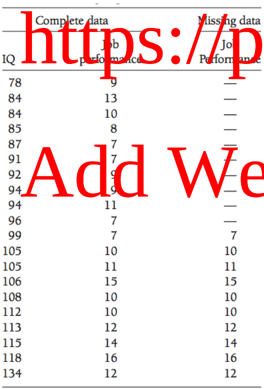
$$\begin{aligned} JP_i &= \hat{\beta}_0 + \hat{\beta}_1(IQ_i) + z_i \\ &= -2.065 + 0.123(IQ_i) + z_i \end{aligned}$$

and $z_i \sim Normal(0, \sigma^2_{JP|IQ})$ where $\sigma^2_{JP|IQ}$ is the residual variance.

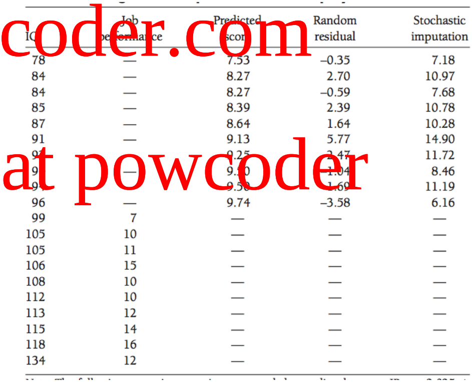
Stochastic regression imputation

Stochastic regression imputation add random residuals to the predicate values generated by standard regression imputation.

- Basic idea: to restore lost variability to the data and effectively eliminate the biases associated with standard regression imputation methods.
- Example:



	Complete data	Missing data
IQ	Job performance	Job Performance
78	9	—
84	13	—
84	10	—
85	8	—
87	7	—
91	7	—
92	9	—
94	9	—
94	11	—
96	7	—
99	7	7
105	10	10
105	11	11
106	15	15
108	10	10
112	10	10
113	12	12
115	14	14
118	16	16
134	12	12



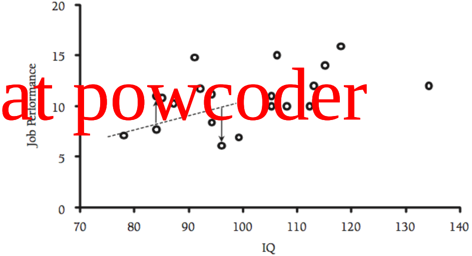
IQ	Job performance	Predicted score	Random residual	Stochastic imputation
78	—	7.53	–0.35	7.18
84	—	8.27	2.70	10.97
84	—	8.27	–0.59	7.68
85	—	8.39	2.39	10.78
87	—	8.64	1.64	10.28
91	—	9.13	5.77	14.90
92	—	9.25	2.47	11.72
94	—	9.50	–1.04	8.46
94	—	9.50	1.69	11.19
96	—	9.74	–3.58	6.16
99	7	—	—	—
105	10	—	—	—
105	11	—	—	—
106	15	—	—	—
108	10	—	—	—
112	10	—	—	—
113	12	—	—	—
115	14	—	—	—
118	16	—	—	—
134	12	—	—	—

Stochastic regression imputation

Stochastic regression imputation add random residuals to the predicate values generated by standard regression imputation.

- Basic idea: to restore lost variability to the data and effectively eliminate the biases associated with standard regression imputation methods.
- Example:

IQ	Job performance	Predicted score	Random residual	Stochastic imputation
78	—	7.53	–0.35	7.18
84	—	8.27	2.70	10.97
84	—	8.27	–0.59	7.68
85	—	8.39	2.39	10.78
87	—	8.64	1.64	10.28
91	—	9.13	5.77	14.90
92	—	9.25	2.47	11.72
94	—	9.50	–1.04	8.46
94	—	9.50	1.69	11.19
96	—	9.74	–3.58	6.16
99	7	—	—	—
105	10	—	—	—
105	11	—	—	—
106	15	—	—	—
108	10	—	—	—
112	10	—	—	—
113	12	—	—	—
115	14	—	—	—
118	16	—	—	—
134	12	—	—	—



Stochastic regression imputation

Stochastic regression imputation add random residuals to the predicate values generated by standard regression imputation.

- Basic idea: to restore lost variablity to the data and effectively eliminate the biases associated with standard regression imputation methods.
- The only procedure in this chapter that gives unbiased parameter estimates under an MAR missing data mechanism.

Imputation with K-Nearest Neighbour

- The idea: use value of the K-Nearest neighbours to impute the missing value.
- Estimate a missing value $y_{i,h}$ in the i-th observation y_i
  - Select K observations whose attribute values are similar to y_i
  - the missing value is estimated as
    - categorial values: the most common values among all neighbours
    - numerical values: the average value is used
- weighted KNNI

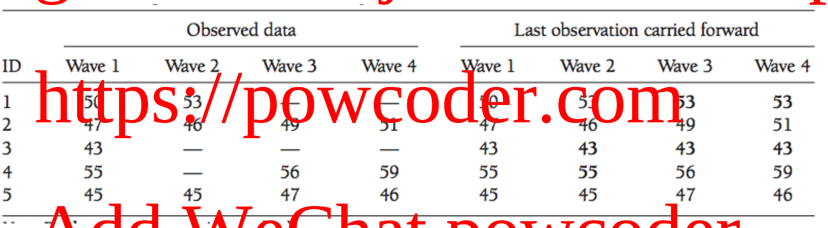
$$y_{\hat{i},h} = \frac{\sum_{j \in I_{Kih}} s_i(y_j) y_{j,h}}{\sum_{j \in I_{Kih}} s_i(y_j)}$$

Other imputation methods

- **Hot-deck imputation**: a collection of techniques that impute the missing values with scores from "similar" respondents.
  - Example: consider a general population survey in which some respondents refuse to disclose their income.
    - classifies respondents into cells based on demographic characteristics such as gender, age, race, and marital status
    - replaces the missing values with a random draw from the income distribution of respondents that shared the same constellation of demographic characteristics as the individual with missing data.

Other imputation methods

- **Last observation carried forward**: specific to longitudinal designs
    - imputes missing repeated measures variables with the observation that immediately precedes dropout



	Observed data				Last observation carried forward			
ID	Wave 1	Wave 2	Wave 3	Wave 4	Wave 1	Wave 2	Wave 3	Wave 4
1	50	53	—	—	50	53	53	53
2	47	46	49	51	47	46	49	51
3	43	—	—	—	43	43	43	43
4	55	—	56	59	55	55	56	59
5	45	45	47	46	45	45	47	46

Evaluate a missing-data method

- **Minimise bias**: Although it is well-known that missing data can introduce bias into parameter estimates, a good method should make that bias as small as possible.
- **Maximise the use of available information**: We want to avoid discarding any data, and we want to use the available data to produce parameter estimates that are efficient (i.e., have minimum sampling variability).
- **Yield good estimates of uncertainty**: We want accurate estimates of standard errors, confidence intervals and p-values.

Summary

- What we discussed
  - Missing value patterns
  - Missing value mechanisms
  - Different methods used to handle missing values
- Acknowledgement: this content of those slides are based on
  - Chapters 1 and 2 in "Applied Missing Data Analysis" by Craig K. Enders
  - "A review of methods for missing data" by Therese D. Pigott
  - 'Dealing with missing data: Key assumptions and methods for applied analysis" by Marina Soley-Bori
  - Chapters 2 and 3 in "Missing data" by Paul D. Allison
- **Assessment 2** released.
  - Due date: Wednesday 3 Oct.