Handling Missing Data in Research: A Practical Guide

Traditional Methods

Waylon Howard

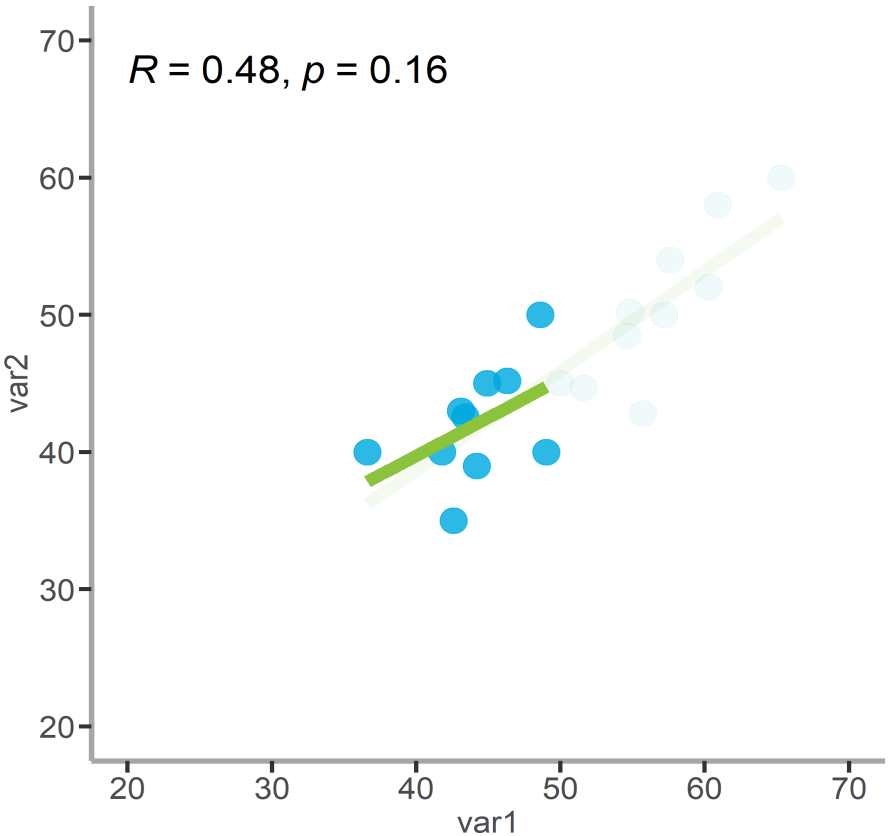
Webinar, November 19, 2024

Listwise (or pairwise) deletion

- Drop cases with missing data
- Default (in virtually every program)
- Requires MCAR
- Power loss (N decrease)

ID	var1	var2
1	36.60	40.00
2	41.80	40.00
3	42.60	35.00
4	43.10	43.00
5	43.40	42.60
6	44.20	39.00
7	44.90	45.00
8	46.30	45.20
9	48.60	50.00
10	49.00	40.00
11	50.00	
12	51.60	
13	54.60	
14	54.80	
15	55.70	
16	57.20	
17	57.60	
18	60.30	
19	60.90	
20	65.30	

Deletion 🤨



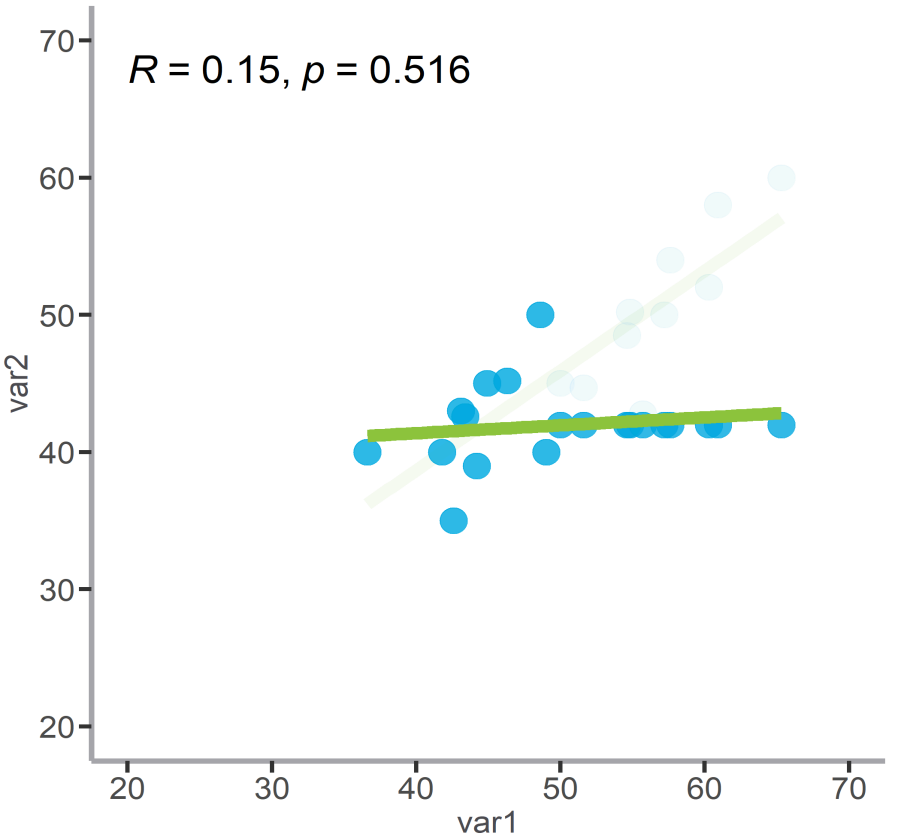
Complete Data Reference: $R = 0.85$, $p < .001$

Mean Imputation

- Replace missing values with the average
- Super simple and seems intuitive
- Can annihilate variability
- Likely the worst option

ID	var1	var2
1	36.60	40.00
2	41.80	40.00
3	42.60	35.00
4	43.10	43.00
5	43.40	42.60
6	44.20	39.00
7	44.90	45.00
8	46.30	45.20
9	48.60	50.00
10	49.00	40.00
11	50.00	41.98
12	51.60	41.98
13	54.60	41.98
14	54.80	41.98
15	55.70	41.98
16	57.20	41.98
17	57.60	41.98
18	60.30	41.98
19	60.90	41.98
20	65.30	41.98

Mean Imputation 😳



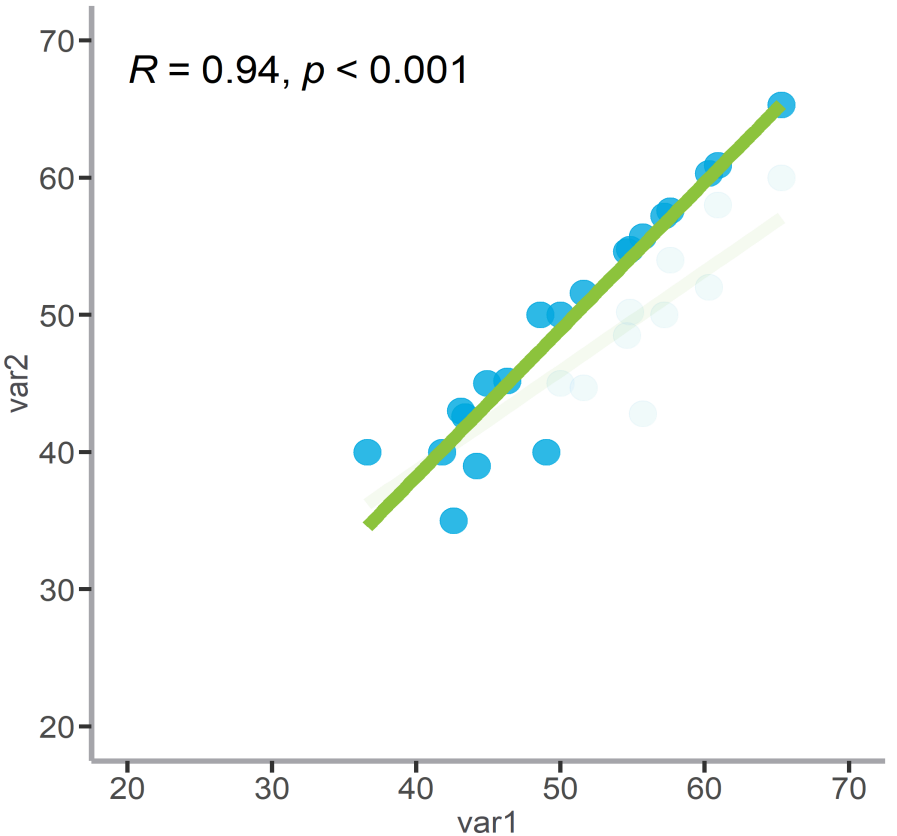
Complete Data Reference: $R = 0.85$, $p < .001$

Last Observation Carried Forward

- Replace missing values with last observed score
- Simple and again seems intuitive
- Can drastically inflate associations

ID	var1	var2
1	36.60	40.00
2	41.80	40.00
3	42.60	35.00
4	43.10	43.00
5	43.40	42.60
6	44.20	39.00
7	44.90	45.00
8	46.30	45.20
9	48.60	50.00
10	49.00	40.00
11	50.00	50.00
12	51.60	51.60
13	54.60	54.60
14	54.80	54.80
15	55.70	55.70
16	57.20	57.20
17	57.60	57.60
18	60.30	60.30
19	60.90	60.90
20	65.30	65.30

LOCF 🙈 🙉 🙊



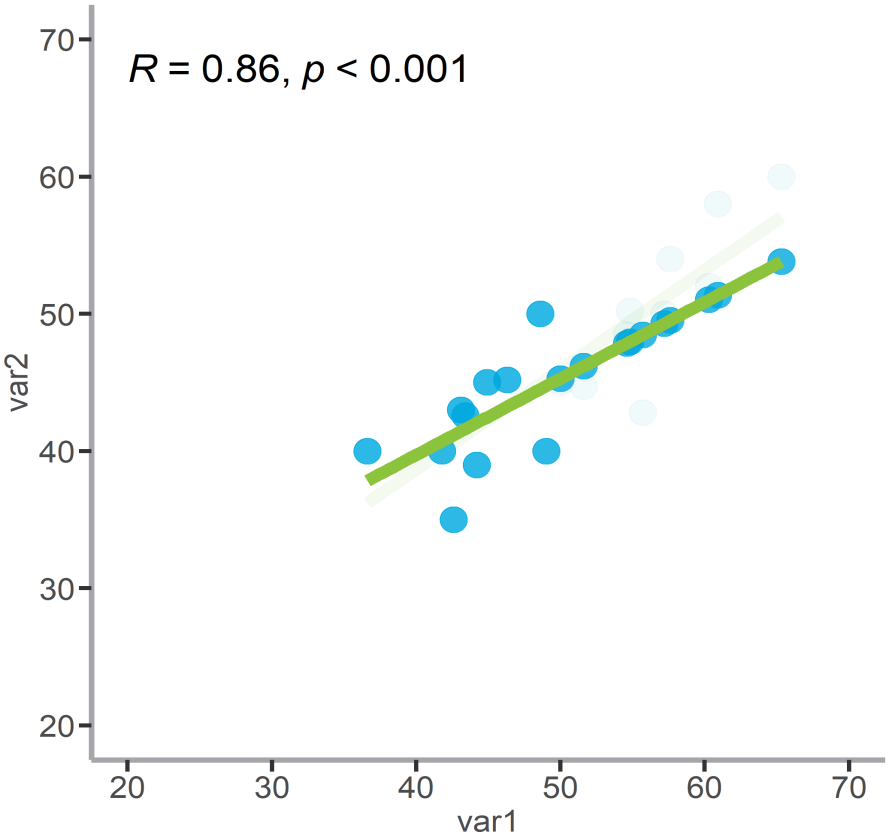
Complete Data Reference: $R = 0.85$, $p < .001$

Single Imputation

- Replace missing values with the predicted score
- Imputing from observed values
- Does not capture variability (regression line)
- Inflates associations

ID	var1	var2
1	36.60	40.00
2	41.80	40.00
3	42.60	35.00
4	43.10	43.00
5	43.40	42.60
6	44.20	39.00
7	44.90	45.00
8	46.30	45.20
9	48.60	50.00
10	49.00	40.00
11	50.00	45.29
12	51.60	46.18
13	54.60	47.85
14	54.80	47.96
15	55.70	48.46
16	57.20	49.30
17	57.60	49.52
18	60.30	51.02
19	60.90	51.36
20	65.30	53.81

One Imputation 🤔



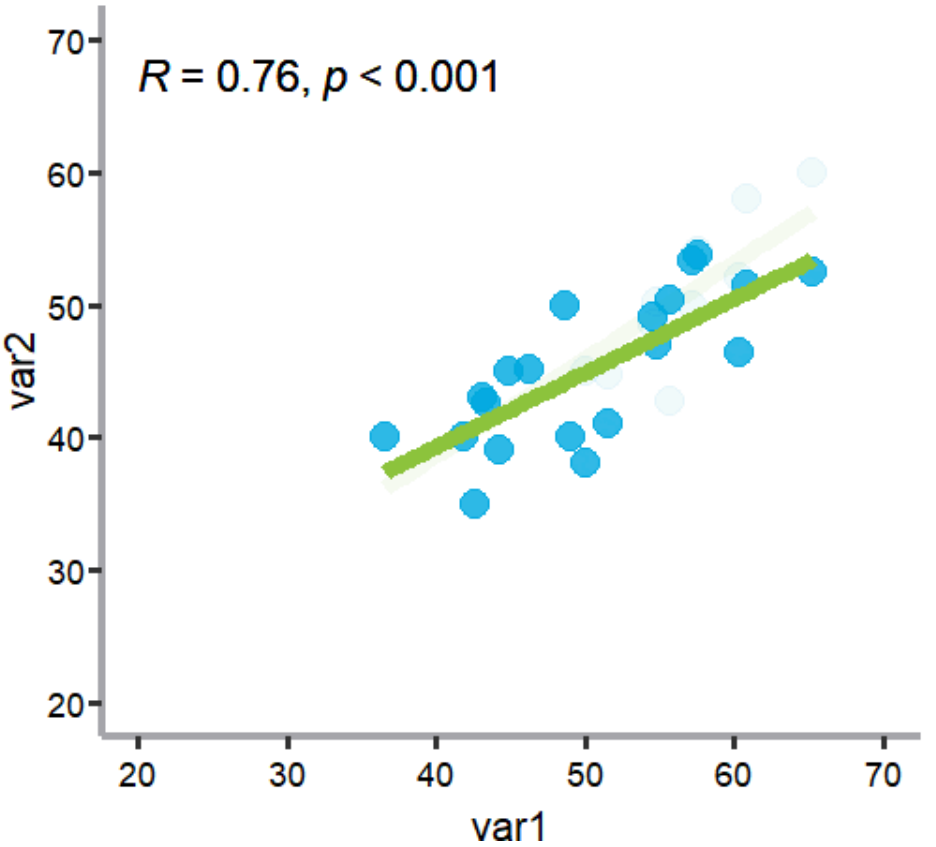
Complete Data Reference: $R = 0.85$, $p < .001$

Stochastic Imputation

- Replace missing values with regression-based estimate and add variability
- Better but does not capture the variability across imputations

Stochastic

ID	var1	var2
1	36.60	40.00
2	41.80	40.00
3	42.60	35.00
4	43.10	43.00
5	43.40	42.60
6	44.20	39.00
7	44.90	45.00
8	46.30	45.20
9	48.60	50.00
10	49.00	40.00
11	50.00	38.05
12	51.60	41.03
13	54.60	49.14
14	54.80	46.93
15	55.70	50.43
16	57.20	53.33
17	57.60	53.70
18	60.30	46.39
19	60.90	51.44
20	65.30	52.45



Complete Data Reference: $R = 0.85$, $p < .001$

Any questions?