Mixed Models with R: Missing Data

ICPSR 2017 at York University

Georges Monette

random@yorku.ca

When does missing data matter?

Easy ‘solution’:

- Don’t use any **record** that has missing data for any variable used in the model: “**complete case analysis**”
- With multilevel or longitudinal data, if a Level-1 variable is missing we only drop the single Level-1 observation or the single occasion.
- If a Level-2 variable is missing, we drop the group.
- What can go wrong with complete case analysis?

An example:

David Reid and Saunia Ahmad at York University do research on couple counselling. The ultimate goal of ***X : treatment*** is to improve ***Y: relationship satisfaction***. But the proximate target is increasing ***W: couple identity (Weness).***

To visualize better, we consider treatment to be a continuous variable, a kind of dose of psychotherapy. Pretend it corresponds to different durations. The hypothesis (and, it seems, reality) is that

$$\mathbf{X} \Rightarrow \mathbf{W} \Rightarrow \mathbf{Y}$$

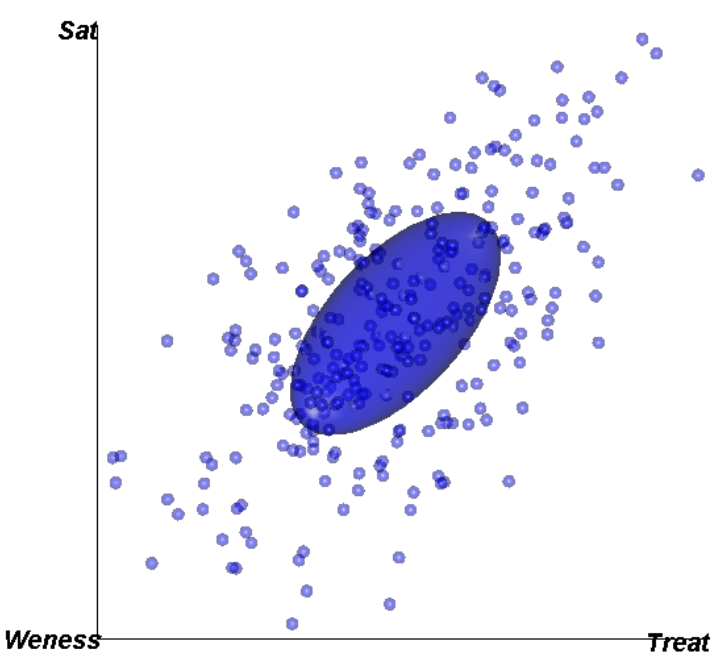
To test **whether X causes Y**, we need to regress Y on X alone: i.e. Y ~ X.

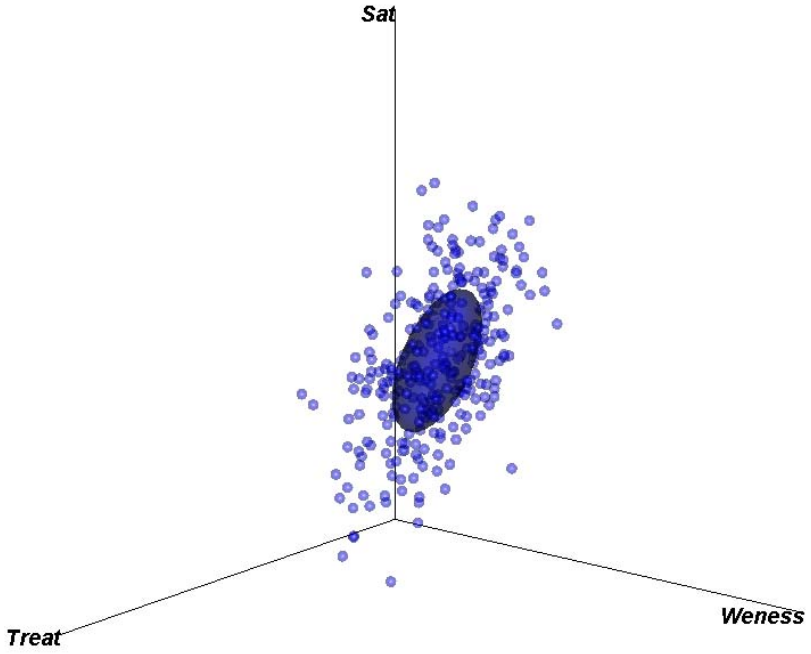
To test **how X causes Y**, we would consider the models Y ~ X + W, W ~ X and Y ~ W.

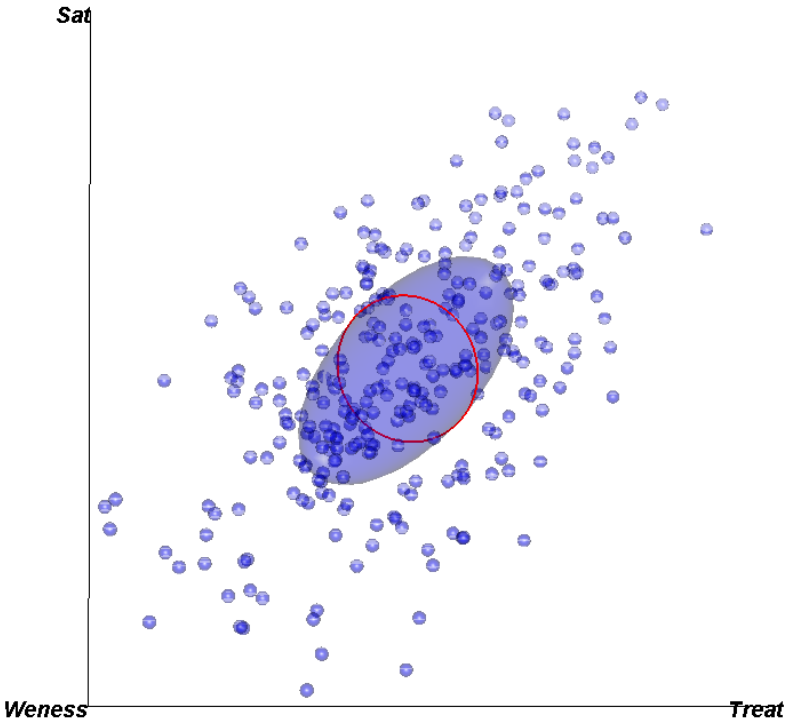
Let's see what happens under different scenarios for missingness.

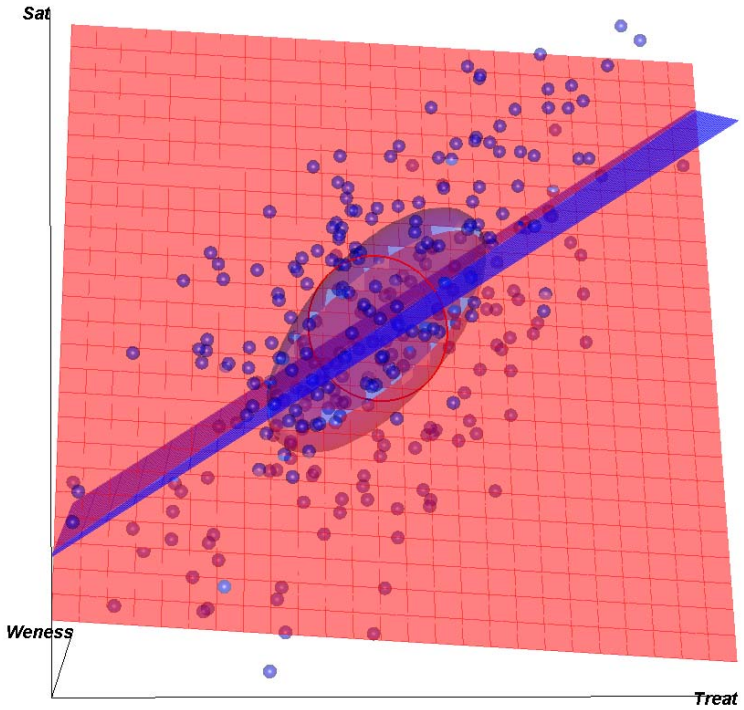
Full data:

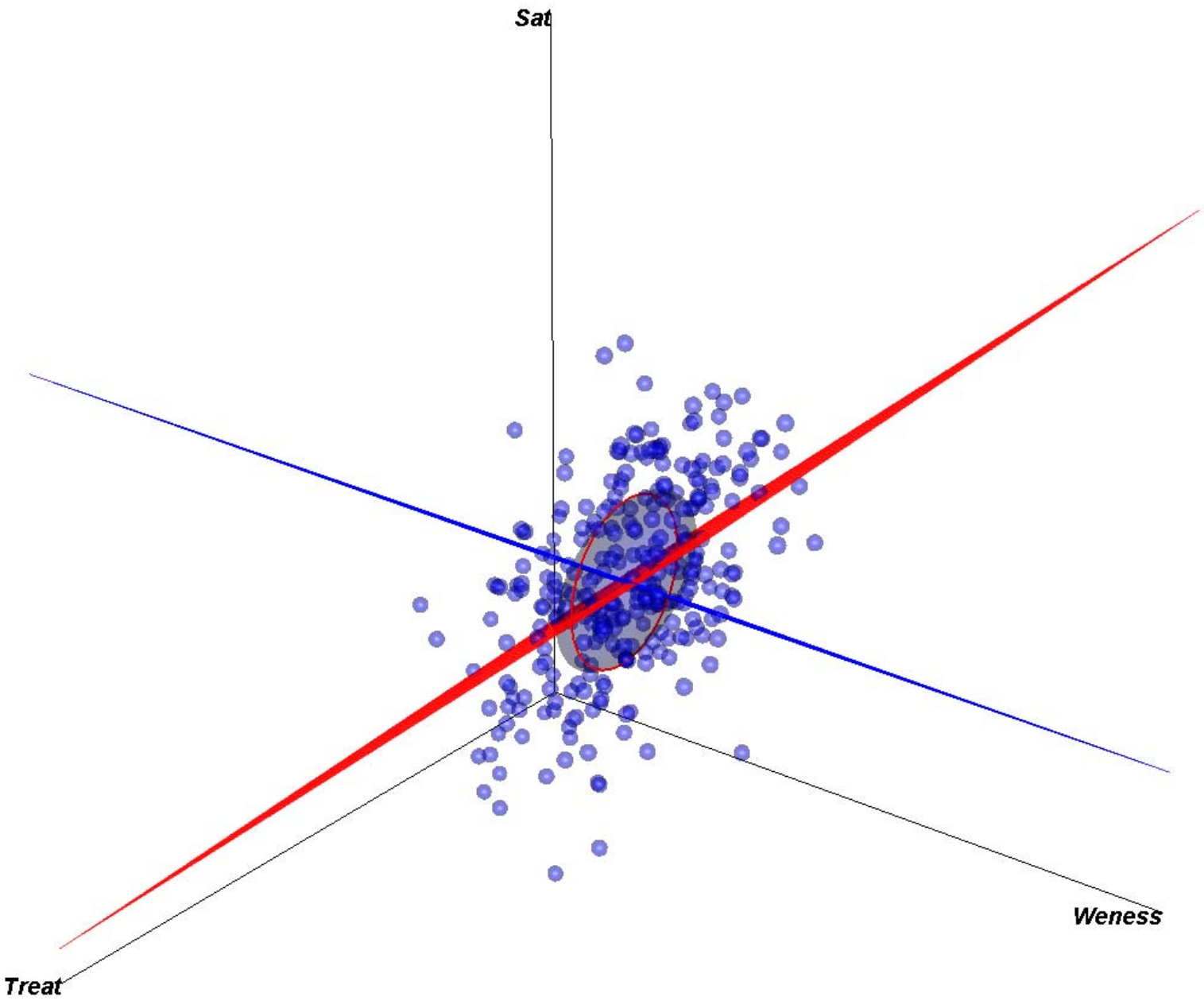
```
> dim(dd)
[1] 300   3
> head(dd)
     Treat   Weness      Sat
1 46.68287 51.07492 57.37784
2 43.26238 35.99149 42.69405
3 57.16086 60.41643 54.41712
4 35.33507 47.78382 56.02111
5 46.92982 52.88721 57.86344
6 59.31408 66.13087 65.04111
```


Sat
Weness
Treat


Sat
Treat
Weness


Sat
Weness
Treat


Sat
Weness
Treat


Sat
Weness
Treat

```
> summary(fitm)

Call:
lm(formula = Sat ~ Treat, data = dd)

Residuals:
     Min       1Q   Median       3Q      Max
-24.0808  -4.4128  -0.1708   5.5025  16.5643

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 20.37433    2.32584    8.76   <2e-16 ***
Treat        0.58652    0.04558   12.87   <2e-16 ***
```

Treat is significant in Y ~ Treat

```
> summary(fitfull)

Call:
lm(formula = Sat ~ Treat + Weness, data = dd)

Residuals:
     Min       1Q   Median       3Q      Max
-18.8999  -4.1144  -0.2531   4.1300  16.8789

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  9.67720    1.91320   5.058 7.43e-07 ***
Treat       -0.05820    0.05582  -1.043    0.298
Weness       0.85301    0.05786  14.744  < 2e-16 ***
```

But not in

Y ~ Treat + Weness

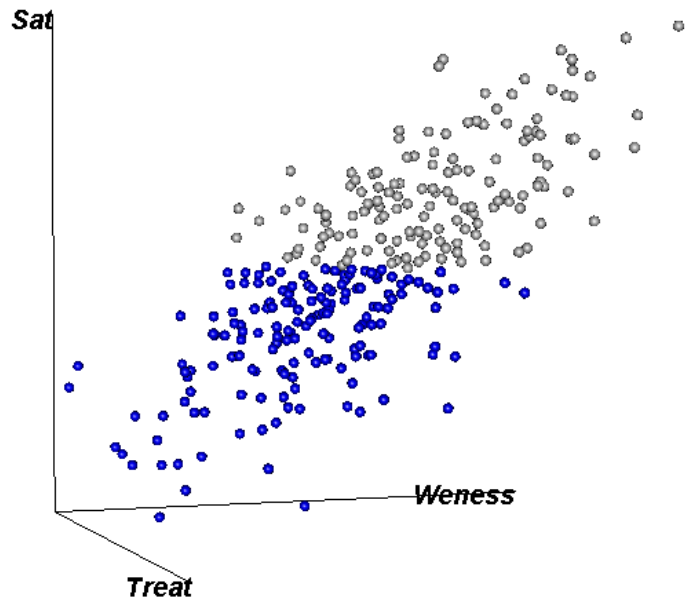
- Let us see what happens when we have missing data for Y

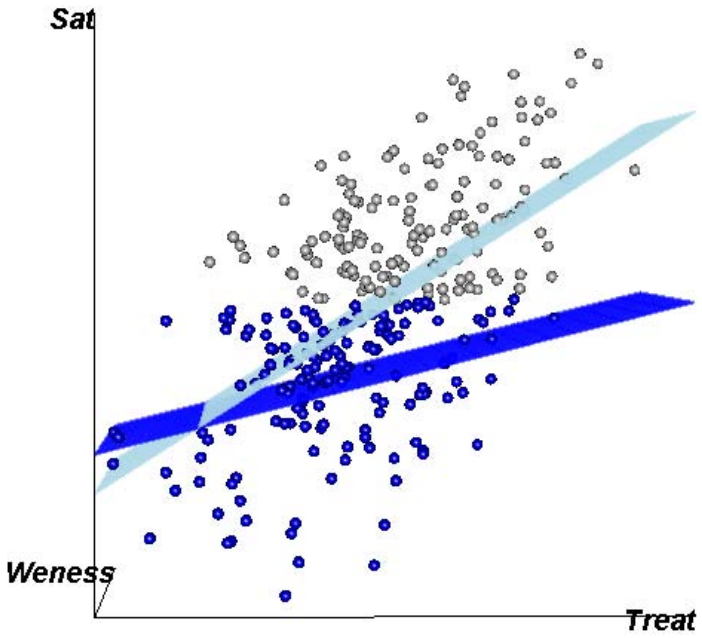
- Three mechanisms:
    1) Missing if Y > 50
    2) Missing if X > 50
    3) Missing if W > 50

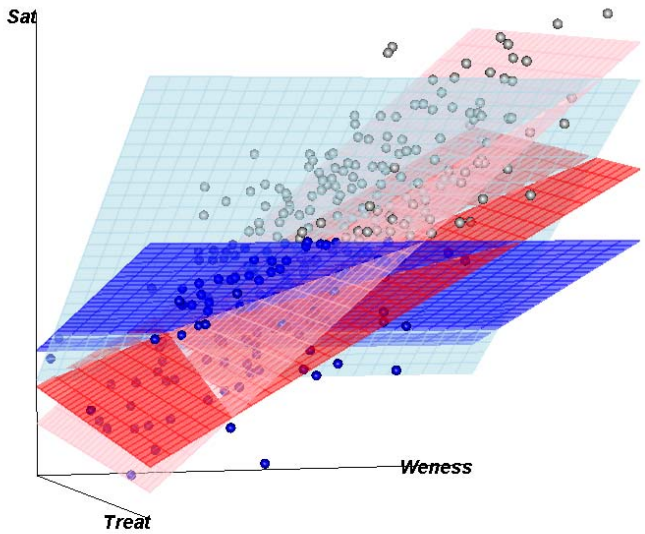
- Consequences for CCA (Complete Case Analysis):
    - What happens if we perform analyses only using complete cases.
      i.e. rows with no missing data.

Missingness due to Y

Y is missing if Y > 50







```
Model: lm(formula = Sat ~ Treat, data = dd)
```

All Data

	Estimate	Std. Error	t value	Pr(>|t|)	
Treat	0.58652	0.04558	12.87	<2e-16	***

Complete Cases (missing if Y > 50)

	Estimate	Std. Error	t value	Pr(>|t|)	
Treat	0.23295	0.05289	4.404	1.98e-05	***

```
Model: lm(formula = Sat ~ Treat + Weness, data = dd)
```

All Data

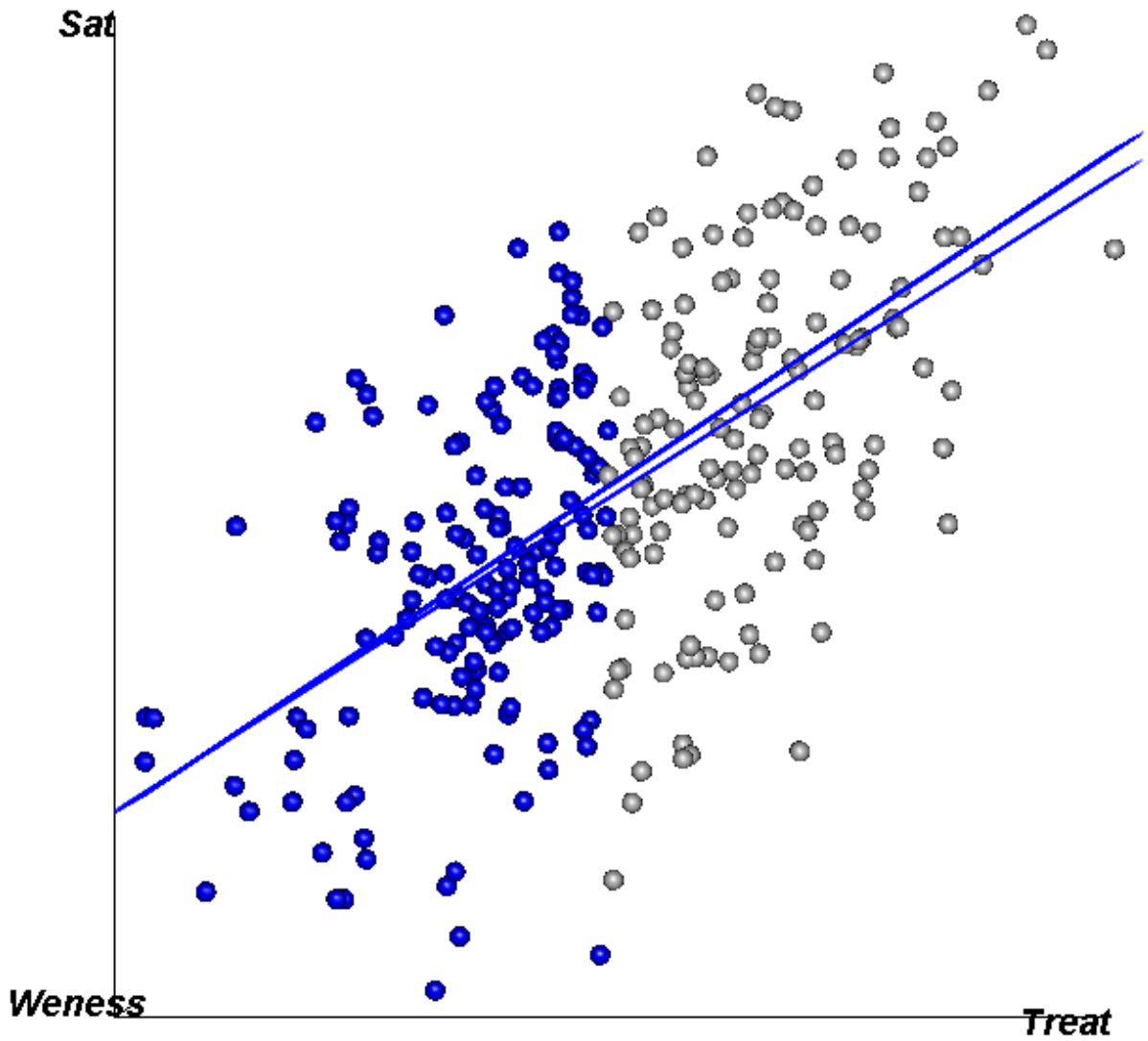
	Estimate	Std. Error	t value	Pr(>|t|)	
Treat	-0.05820	0.05582	-1.043	0.298	
Weness	0.85301	0.05786	14.744	< 2e-16	***

Complete Cases (missing if Y > 50)

	Estimate	Std. Error	t value	Pr(>|t|)	
Treat	-0.08649	0.06186	-1.398	0.164	
Weness	0.55795	0.07367	7.573	3.21e-12	***

So missingness affects **both** regressions.

Missingness due to Treatment: Y missing if X > 50



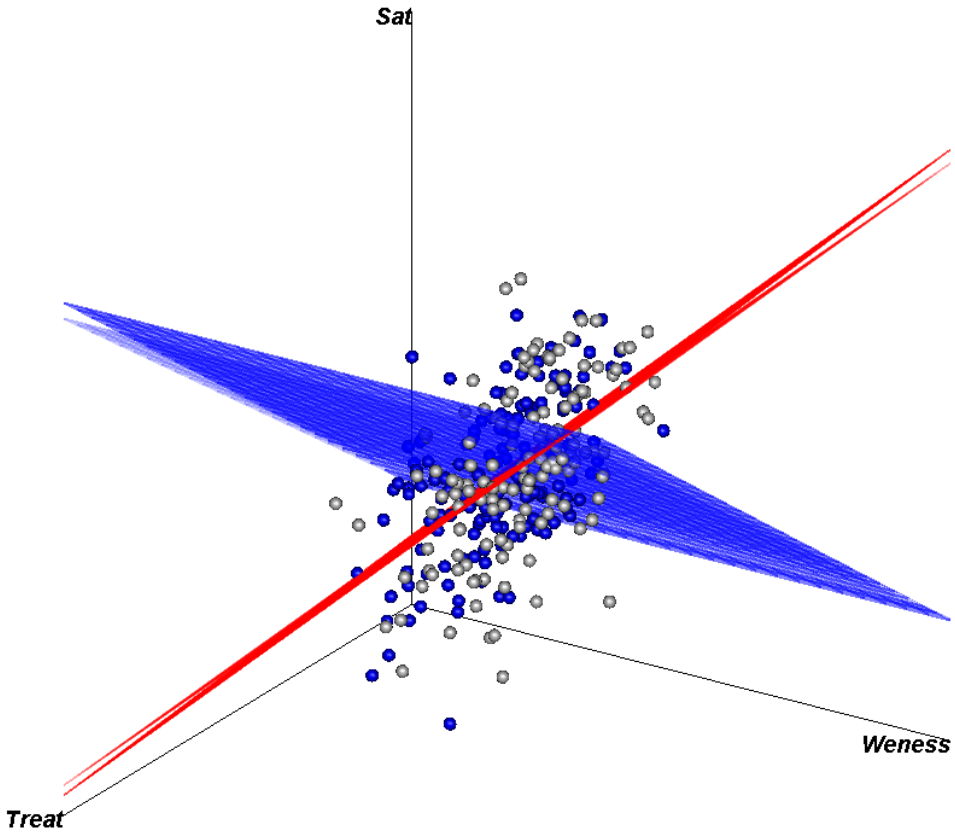
```
Model: lm(formula = Sat ~ Treat, data = dd)

All Data:
            Estimate Std. Error t value Pr(>|t|)
Treat        0.58652    0.04558   12.87   <2e-16 ***

Complete Cases (missing if X > 50):
            Estimate Std. Error t value Pr(>|t|)
Treat          0.6130     0.1071   5.721 5.30e-08 ***
```

Note: Coefficient did not change 'much' but SE increased by a factor of 2.35 – approximately what we expected because the SD of X was reduced by a factor of about .6 and n was reduced by 1/2.

The effect on the SE of $\hat{\beta}_X$ is to increase by a factor of $\frac{1}{.6\times\sqrt{1/2}} = 2.36$



Model: lm(formula = Sat ~ Treat + Weness, data = dd)

All Data:

	Estimate	Std. Error	t value	Pr(>|t|)	
Treat	-0.05820	0.05582	-1.043	0.298	
Weness	0.85301	0.05786	14.744	< 2e-16	***

Complete Cases (missing if X > 50):

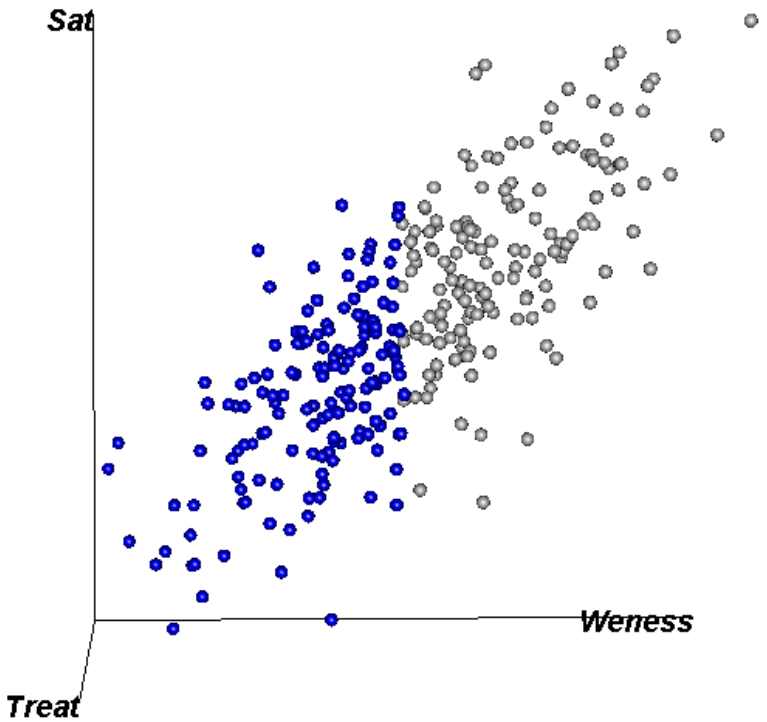
Treat	-0.07726	0.10074	-0.767	0.4443	
Weness	0.87185	0.07775	11.214	<2e-16	***

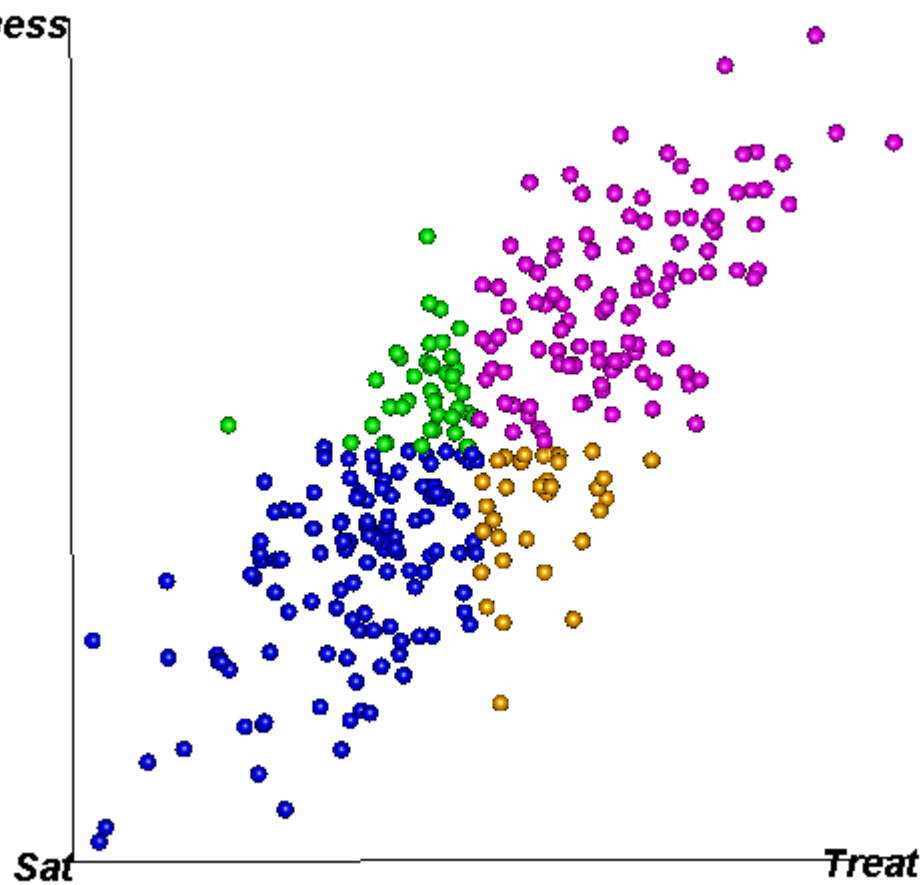
Recap so far:

If missingness is due to Y, both models go bad.

If missingness is due to X, both models are unbiased – although SEs are larger, as expected.

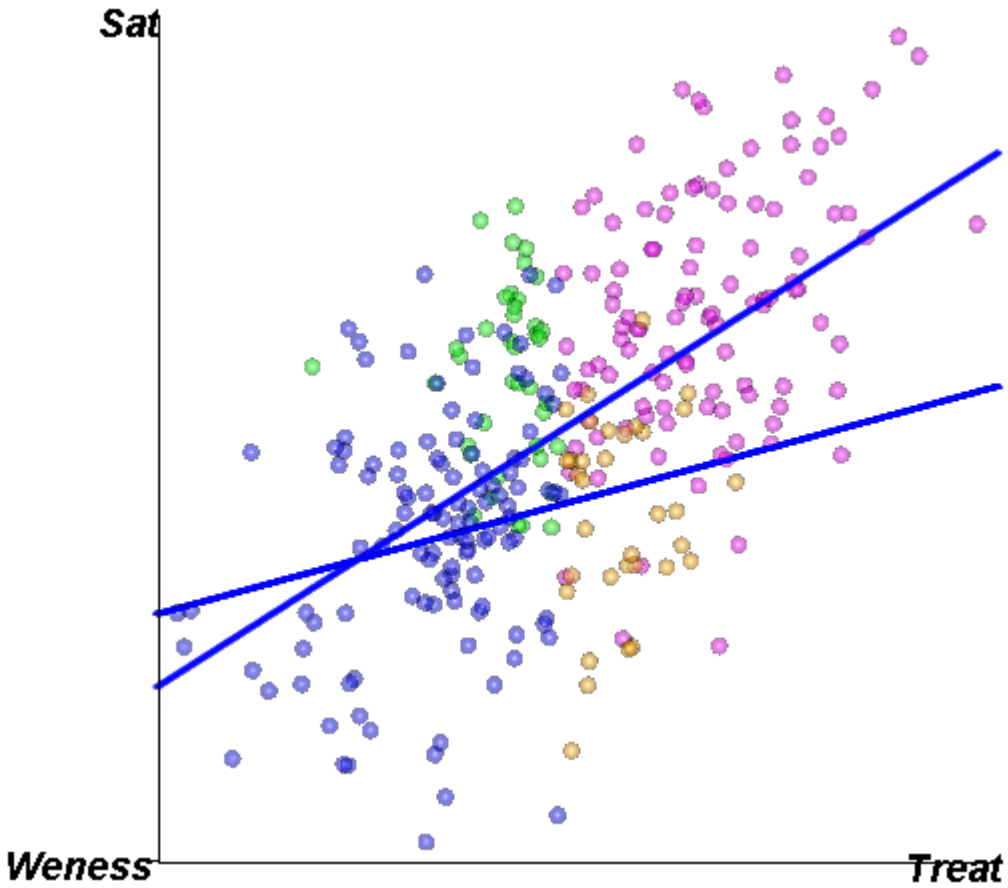
Missingness due to Mediator: Y missing if Weness > 50



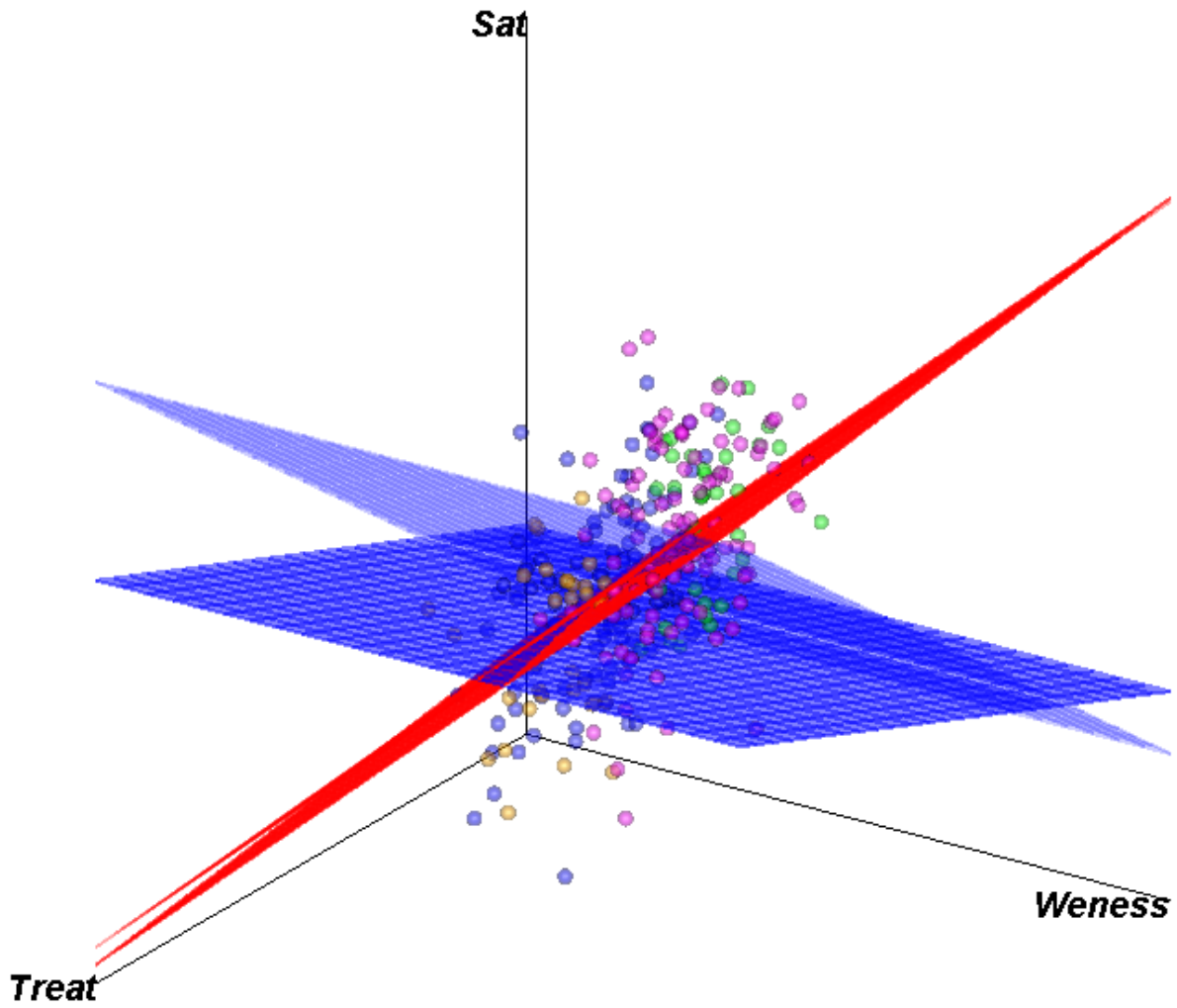


The green points are present under ‘Treat missingness’ but absent under ‘Weness missingness’. Conversely for the orange points. So going from Treat-missingnes to Weness-missingness, we remove the green points and add the orange points.

The effect is to flatten the regression on Treat.



But the relationship of Sat with both Treat and Weness is not seriously affected:



Model: lm(formula = Sat ~ Treat, data = dd)

All Data:

```
            Estimate Std. Error t value Pr(>|t|)
Treat        0.58652    0.04558   12.87   <2e-16 ***
```

Complete Cases (missing if Weness > 50)

```
            Estimate Std. Error t value Pr(>|t|)
Treat        0.25010    0.07652   3.268  0.00134 **
```

Model: lm(formula = Sat ~ Treat + Weness, data = dd)

All Data:

```
            Estimate Std. Error t value Pr(>|t|)
Treat       -0.05820    0.05582  -1.043    0.298
Weness       0.85301    0.05786  14.744  < 2e-16 ***
```

Complete Cases (missing if Weness > 50)

```
            Estimate Std. Error t value Pr(>|t|)
Treat       -0.10097    0.07681  -1.315  0.19068
Weness       0.83638    0.10222   8.182 1.17e-13 ***
```

Classifying missingness

MCAR = Missing Completely at Random (given fixed non-missing predictors). This means that knowing the residual from a regression on fixed non-missing predictors does not help to predict missingness.

MAR = Missing at Random: Probability that a row or case has missing data depends on missing values only through non-missing data. i.e. Missing values contain no additional information on probability that data is missing.

MNAR = none of the above: missing values contain information that is not already contained in the non-missing data.

Classical example: Consider therapy for depression. You want to estimate the rate of improvement over time. Patients are examined and scored on Y = Depression weekly until they drop out:

MAR: They stop coming AFTER they have achieved a sufficiently low Y:

Missingness depends on a value of Y that has been observed.

MNAR: They stop coming on the first occasion when they feel they have recovered.

Missingness depends on the unobserved value of Y.

For regression models: $Y = X\beta + \varepsilon$, we think of X as fixed and we think of the distribution of Y as dependent on X, i.e. we can think of ε as the random component of the model.

MCAR: Conditionally on fixed components of model, missingness is independent of *Y*.

MAR: Conditionally on fixed components of model, missingness is independent of unobserved components of *Y*.

Key point:

Whether missingness is MCAR or MAR depends on missingness mechanism ***and*** on model.

not MAR

not MAR

		Model	
		Y ~ X	Y ~ X + W
Missingness mechanism	Y > 50	~~not MCAR~~	~~not MCAR~~
	X > 50	MCAR	MCAR
	W > 50	**not MCAR but could be MAR with right model**	MCAR

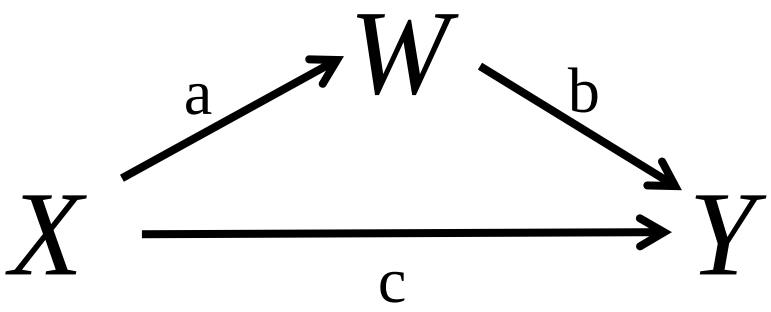
If missingness is due to X, then X must be in the model to achieve MCAR.

If missingness is MCAR for a model, then Complete Case Analysis gives unbiased estimates of coefficients – although CCA might not make best use of the available information.

What can we do with Y~X when Y is missing if W > 50?

In this case missingness is **MAR** (missing at random) if we could think of (Y,W) as random and note that the probability of a case having missing data (for Y) is a function of observed data (W). We could create a FIML (Full Information Maximum Likelihood Model) for (Y_{obs},W | X).

In the usual display of coefficients for a mediation model:



the coefficients a and c are estimated in the regression of (Y,W) on X and the coefficient b is estimated from the variance-covariance estimate of (Y,W) conditional on X. Finally the 'total effect' of X on Y is estimated with $c + ab$.

Example where LME works with MAR:

Simulation exercise:

- Estimate Growth as a function of age
- Subjects drop out *after* first achieving a threshold MAR
- Subjects drop out *before* first achieving a threshold NMAR
- Population average trajectory: 10 + 3 x age
- Observations on N subjects at ages 1 to 20
- Suppose variance matrix for intercept and slope is: $\begin{bmatrix} 1 & 0 \\ 0 & .2 \end{bmatrix}$

  and within-occasion σ is 1.

  1) Generate complete data
  2) Set as missing any y > 40; analyze
  3) Set as missing any ys after the first y equal to 40; analyze
  4) Compare the two analyses. How close do they get to estimating the 'true' population parameters: intercept = 30 and coefficient of age = 3.

The Magic of Multiple Imputation:

If you have missing data, don't throw away the good data that goes with your missing data (i.e. CCA), just make up ('impute') the missing data!

When can this work? If the non-missing data can give you a reasonable guess of the missing data. I.E. the missingness is MAR given all available data – not just the data in analysis model.

Q: But guesses have less variability than real random data – a predicted value generally has less variability than what it predicts. So imputed data might not look like real data.

A: Add random errors to make the imputed data look like real data.

Q: But this is crazy! If I just make up the data, how do I know my results aren't just a random consequence of the data I made up?

A: Do it many times and you will be able to tell how much making up the data contributed to the variability in your results.

The main advantage of multiple imputation: **It separates taking care of missingness from doing the analysis.** You can take care of missingness with imputation and once that is done, do your analysis without looking back.

Important implications for research practice: You can get closer to MAR by adding more variables, e.g. proxies for variable of primary interest. Proxies are hard to include in an analytic model but a cinch to add in an imputation model. So it's worthwhile getting those additional variables.

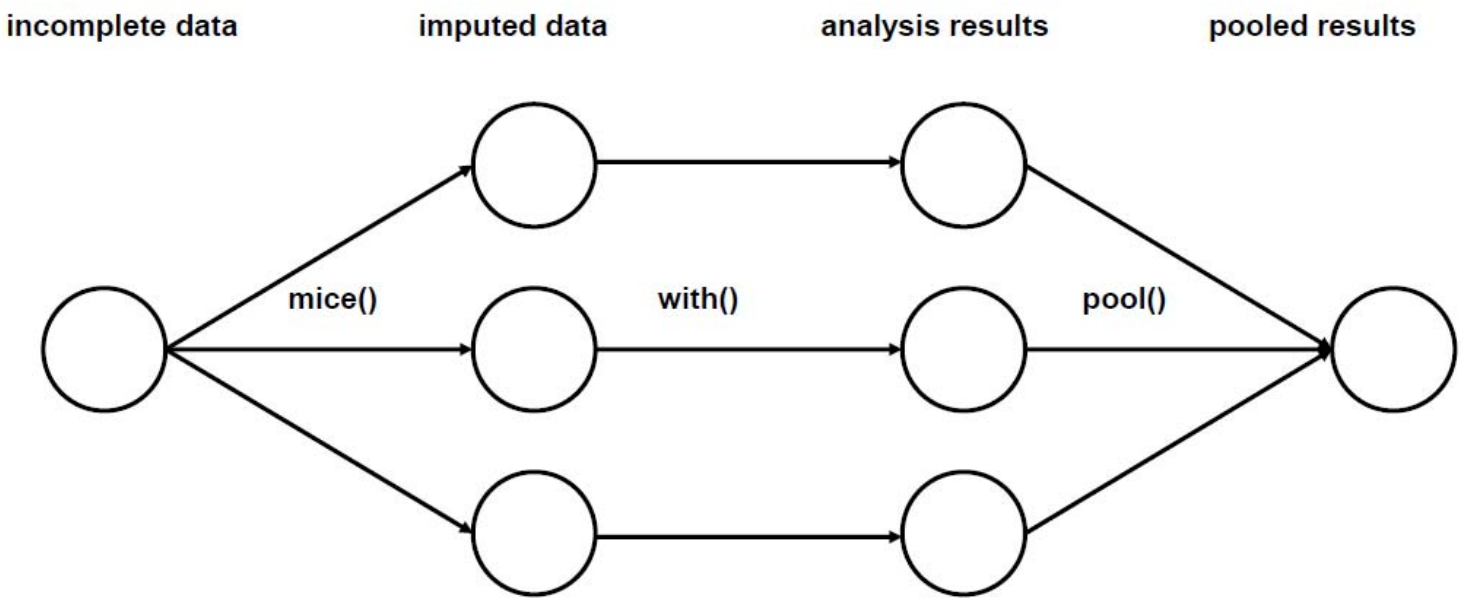
Missing data imputation is a **predictive** problem – thus much easier than a causal problem. Easier to automate, variables can be selected on fit.

A note to statistical consultants: If a client says they have no missing data, that's almost a sure sign that they do. They just tried to do you a favour by throwing away cases with missing data. Suggest that they should get it back!

Summarizing the Magic of Multiple Imputation

- The guessing (imputation) model can use all available variables, whether they are appropriate or not for the *analysis model* (e.g. mediators that must be excluded in a model for causal estimation)
- Having imputed you can use the data for many models.
- You can usefully gather information that would have been useless in an *analysis model*. E.g. proxies for a variable that might have missing data: you can use the proxies to impute but they would be a collinear embarrassment in a model for analysis.
- Caution: should use all *analysis variables* in the *imputation model*.
- Overfitting is not a major issue except to the extent that the distribution of predictor variables in complete data is very different from that of predictor variables where data is missing. Bayesian methods are used to obtain predictive models.

Overview of the Process of Multiple Imputation:



Details of imputation:

Uses MCMC: Steps:

0) Make up missing data somehow – perhaps just use model fitted on non-missing data.
1) Fits imputation model for each Z with missing data on all other variables.
2) Use a Bayesian posterior to generate random parameters for each imputation model.
3) Use random parameter imputation models to generated predicted imputed values for missing data + random error. If more than one variable has missing values, cycle around replacing previous imputed values with new as you impute each variable in turn.
4) If you have iterated long enough to have ‘converged’, use last imputed values as a set of imputed values. Otherwise go back to 1 using newly imputed and original data.

MCMC magic: If you iterate long enough, and the estimation/imputation process doesn't get stuck, you will have imputed values that are a random sample of the missing data given the observed data.

In practice we want at least m = 5 independent sets of imputed values. So we run the process above m times – in parallel (each separate process is called a chain) – and we compare the eventual independence of the chains from their starting position and from each other to assess convergence.

If there is more than one variable with missing data, this process is called 'chained equations' or 'fully conditional specification' FCS. It works with same magic as MCMC.

"Ideally" you would like to generate missing values from the joint distribution for missing variables, but there are too many possibilities. FCS just estimates one imputation model for each individual variable from all the others. How does this work when there is more than one value missing? The process cycles through: make a guess for one variable, then use that in the model to guess the other, etc.

Do Lab on Missing Data.

Question 14 of Andrew Gelman's final exam for Design and Analysis of Sample Surveys

Posted by Andrew on 24 May 2012, 5:00 pm

14. A public health survey of elderly Americans includes many questions, including "How many hours per week did you exercise in your most active years as a young adult?" and also several questions about current mobility and health status. Response rates are high for the questions about recent activities and status, but there is a lot of nonresponse for the question on past activity. You are considering imputing the missing values on the question, "How many hours per week did you exercise in your most active years as a young adult?" Which of the following statements are basically correct? (Indicate all that apply.)

(a) If done reasonably well, imputation is preferred to available-case and complete-case analysis.

(b) If you do impute, you should also present the available-case and complete-case analysis and analyze how the imputed estimates differ.

(c) It is OK to include current health status variables as predictors in a model imputing past activities: anything that adds information is good when imputing.

(d) It is probably not a good idea to include current health status variables as predictors in a model imputing past activities: current health is possibly influenced by past activities, and including a casual outcome can bias estimates of a treatment variable.

(e) If you fit a regression model and impute your best prediction for each person (rather than imputing random draws from the predictive distribution), you can have problems because you will be more likely to impute extreme values.

(f) It is a good idea to fit a logistic regression predicting response/nonresponse to the question of interest as a way to look for systematic differences between respondents and nonrespondents on this question.

Reference:

van Buren, S and Groothuis-Oudshoorn, K. (2011) “mice: Multivariate Imputation by Chained Equations in R”, Journal of Statistical Software, 45(3) at http://www.jstatsoft.org/

Question 14 of Andrew Gelman's final exam for Design and Analysis of Sample Surveys

(a) If done reasonably well, imputation is preferred to available-case and complete-case analysis.

(b) If you do impute, you should also present the available-case and complete-case analysis and analyze how the imputed estimates differ.

(c) It is OK to include current health status variables as predictors in a model imputing past activities: anything that adds information is good when imputing.

(d) It is probably not a good idea to include current health status variables as predictors in a model imputing past activities: current health is possibly influenced by past activities, and including a casual outcome can bias estimates of a treatment variable.

(e) If you fit a regression model and impute your best prediction for each person (rather than imputing random draws from the predictive distribution), you can have problems because you will be more likely to impute extreme values.

(f) It is a good idea to fit a logistic regression predicting response/nonresponse to the question of interest as a way to look for systematic differences between respondents and nonrespondents on this question.

Solution: a, b, c, f. Not d (for imputation you want a predictive model not a causal model) and not e (if you impute the best prediction for each case you will understate the variation and be less, not more, likely to impute extreme values).