Improving upon complete case analysis when covariates are missing not at random

Jonathan Bartlett

London School of Hygiene and Tropical Medicine
www.missingdata.org.uk

28th June 2013
Centre for Statistical Methodology, LSHTM

Acknowledgements

This is joint work with:

- James Carpenter (LSHTM)
- Kate Tilling (University of Bristol)
- Stijn Vansteelandt (Ghent University, Belgium)

Support for this work from:

- UK Economic and Social Research Council, RES-189-25-0103 (JB and JC)
- Medical Research Councils, G0900724 (JB, JC, and KT)
- Interuniversity Attraction Poles Programme, P7/06 (SV)

Outline

Outline

Complete case analysis

- Suppose we have some missing data.
- We have a particular analysis we want to perform, for which some variables involved have missing values.
- In a complete case (or complete records) analysis (CCA), we ignore incomplete cases, and run our analysis on the complete ones.
- Obviously estimates are less precise (than a full data analysis).
- But are the estimates biased?

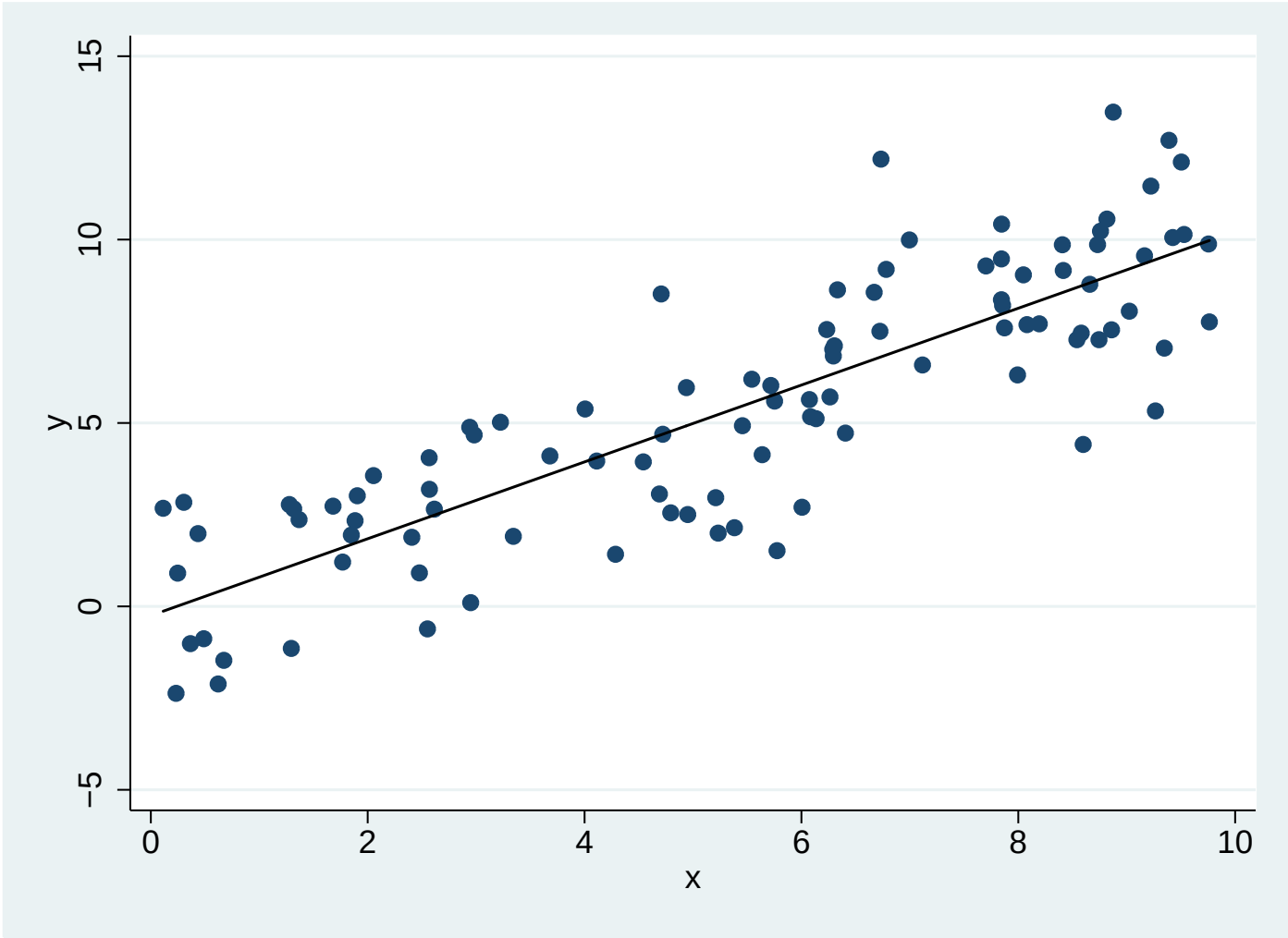
When is complete case analysis valid?

- It is sometimes stated that CCA is only unbiased if data are missing completely at random (MCAR).
- Suppose we are just interested in estimating the mean of a variable Y, i.e. $E(Y)$.
- Let R denote whether Y is observed ($R = 1$) or missing ($R = 0$).
- Here CCA is unbiased if Y and R are independent, i.e. $Y \perp R$.
- If there are no other variables, this means data are MCAR.

When is complete case analysis valid?

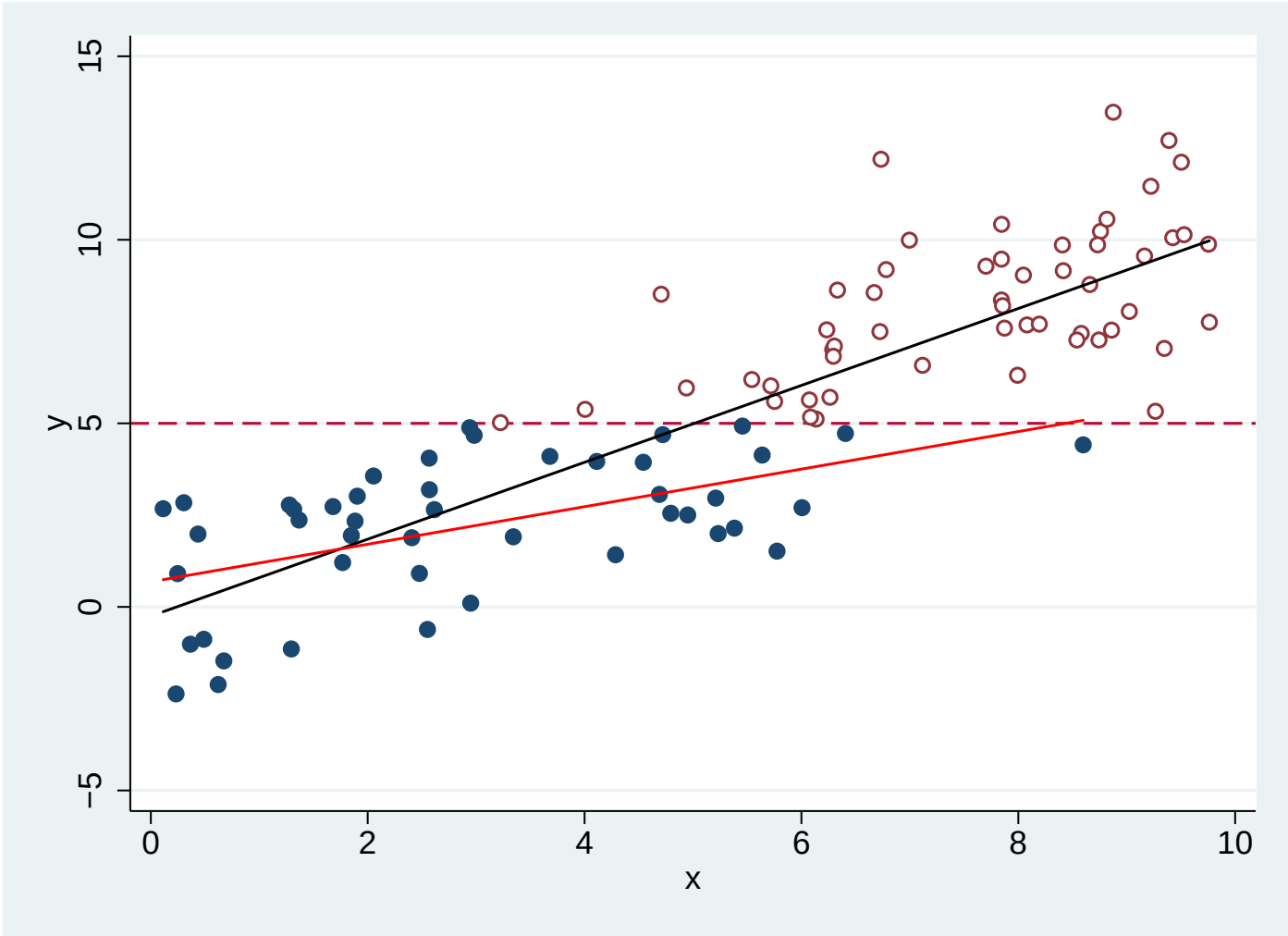
- Now suppose our analysis consists of fitting a regression model, with outcome Y, covariates X.
- We are interested in some aspect of the conditional distribution of $Y|X$, e.g. regression coefficients β.
- Missing values occur either in Y, or X, or maybe both.
- Of course if data are MCAR, CCA is ok, because the complete cases are still a random sample from the population.

Simple linear regression – full data



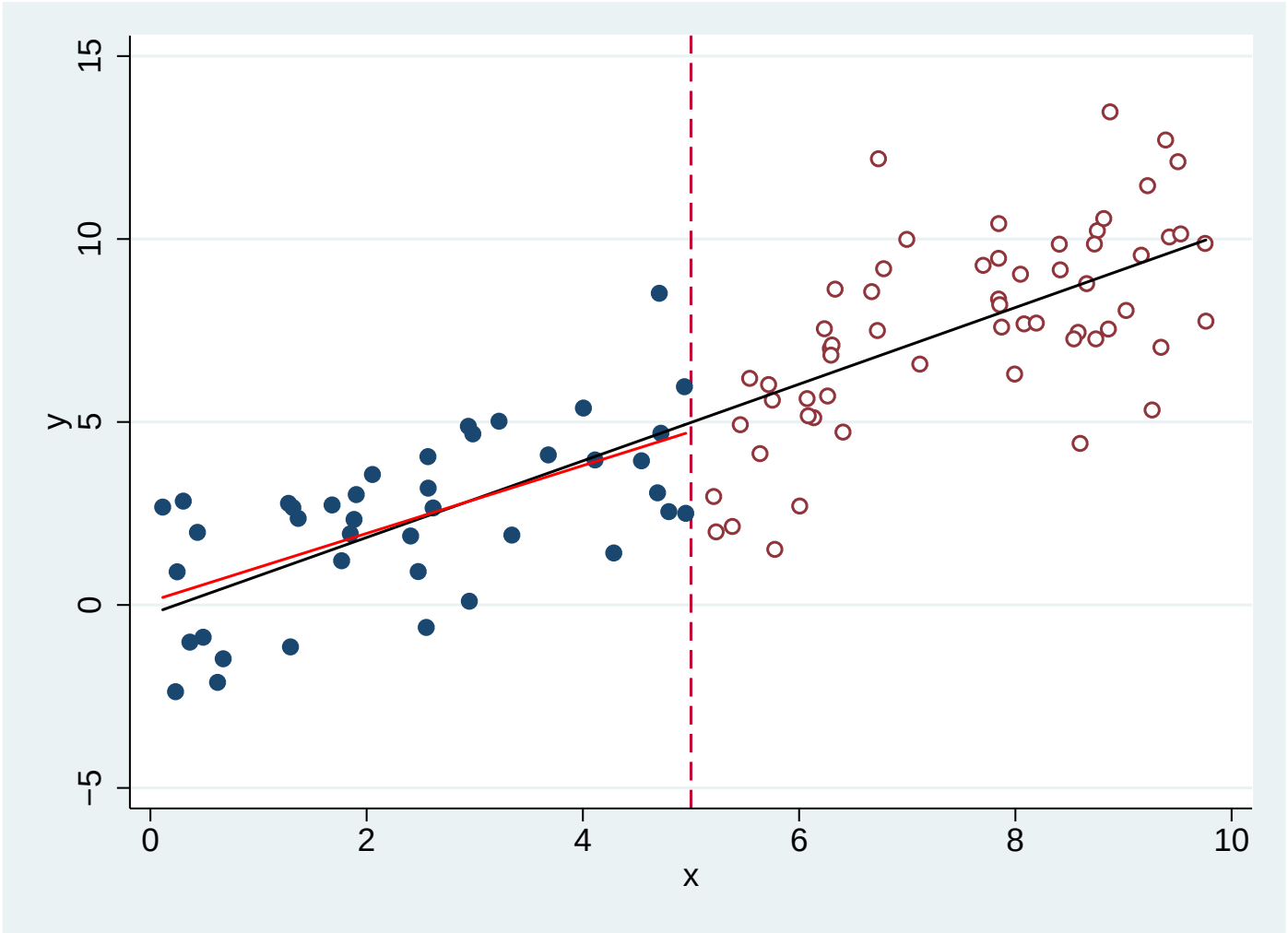
With full data, estimates are unbiased

Missingness dependent on Y



With missingness dependent on Y, we get bias

Missingness dependent on X



With missingness dependent on X, we don't get bias

Covariate dependent missingness

- Let R denote whether a subject is a complete case ($R = 1$) or not ($R = 0$).
- Suppose missingness depends on X, but given X, is independent of Y, i.e. $R \perp\!\!\!\perp Y|X$.
- Then CCA is unbiased for estimation of parameters involved in $Y|X$:

$$
\begin{aligned}
f(Y|X, R=1) = \frac{f(Y, X, R=1)}{f(X, R=1)} &= \frac{f(R=1|X, Y)f(X, Y)}{f(R=1|X)f(X)} \\
&= \frac{f(R=1|X)f(X, Y)}{f(R=1|X)f(X)} \\
&= f(Y|X)
\end{aligned}
$$

i.e. the conditional distribution $Y|X$ in the complete cases is the same as in the population

Covariate dependent missingness and Rubin's taxonomy

- Let us call the assumption that $R \perp\!\!\!\perp Y|X$ the CCA assumption.
- We have not specified in which variables missingness occurs.
- If missingness only occurs in Y,
  the CCA assumption $\equiv$ missing at random (MAR).
- Suppose we divide the covariates into X and Z, where Z is always observed, and X is sometimes missing (denoted by $R = 1$).
- If missingness is dependent on Z only, so $R \perp\!\!\!\perp (Y, X)|Z$, then data are MAR.
- But if missingness is dependent on X, or jointly on X and Z, then data are missing not at random (MNAR).

CCA validity – summary

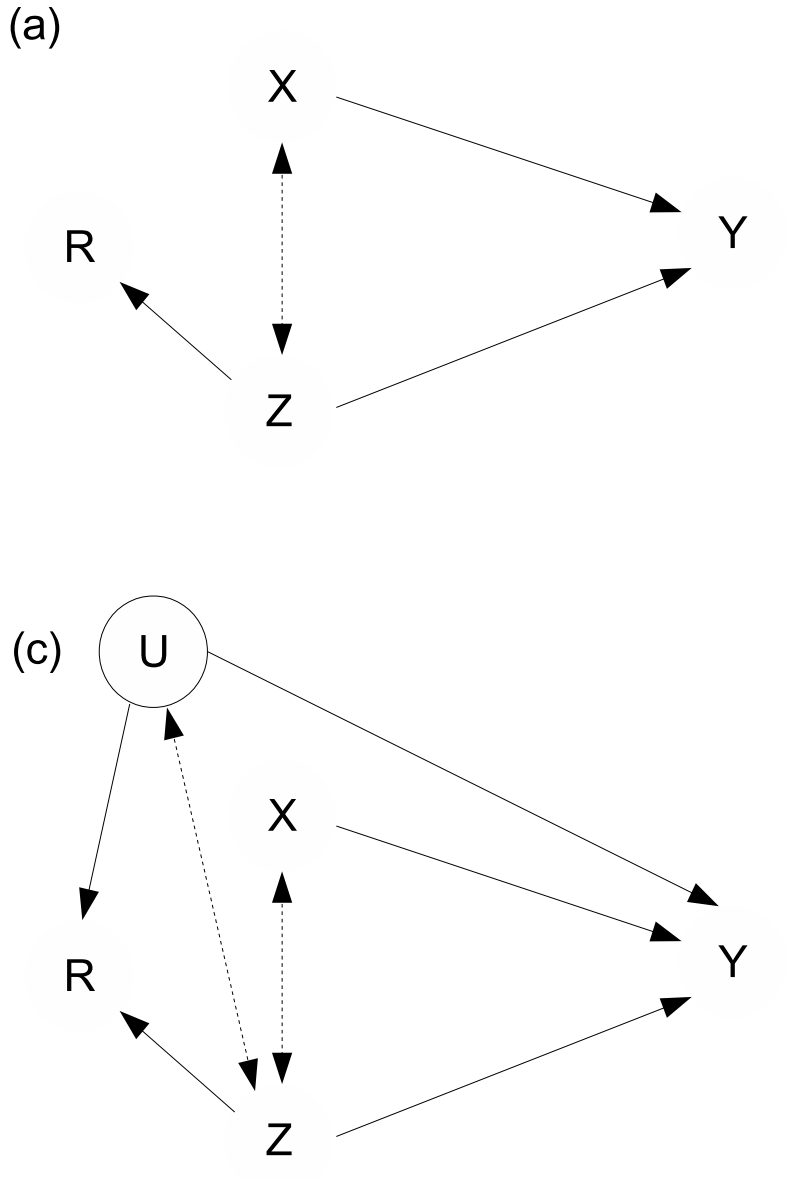
- CCA is always unbiased if data are MCAR.
- CCA can also be unbiased under certain MAR mechanisms.
- CCA can also be unbiased under certain MNAR mechanisms.
- The key is that misssingness is (conditionally) independent of outcome Y.
- For more on this, see [1, 2].

Outline

Covariate dependent missingness

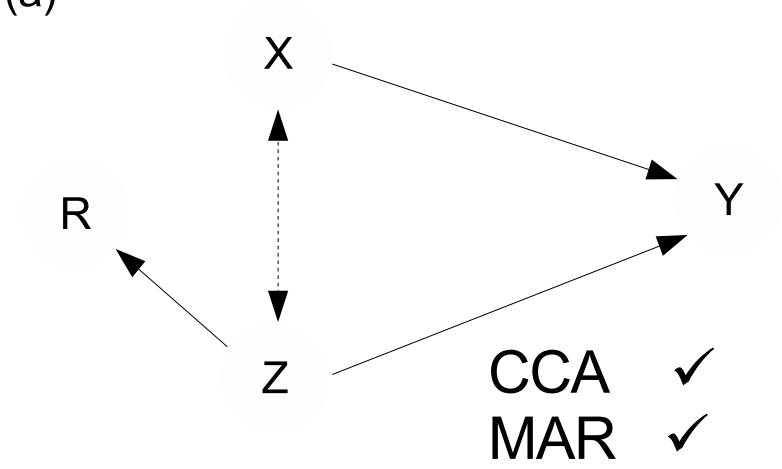
- In some settings it may be plausible that the CCA assumption (covariate dependent missingness), holds.
- Often our covariates are measured at baseline, and the outcome Y is measured later in time.
- In this case, it may be plausible that missingess is independent of the future outcome, conditional on baseline covariates.

Covariate dependent missingness

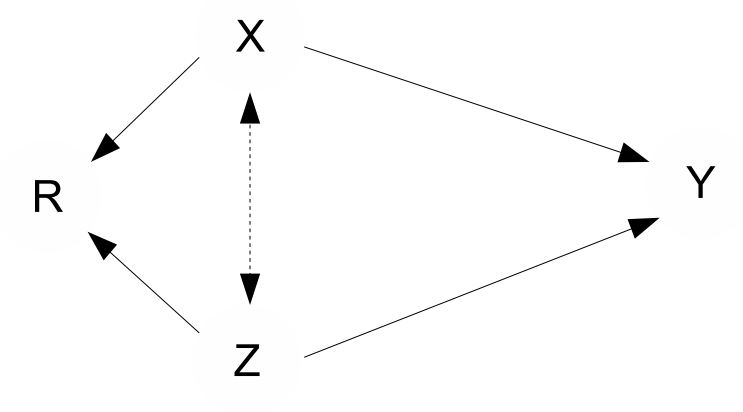


Covariate dependent missingness

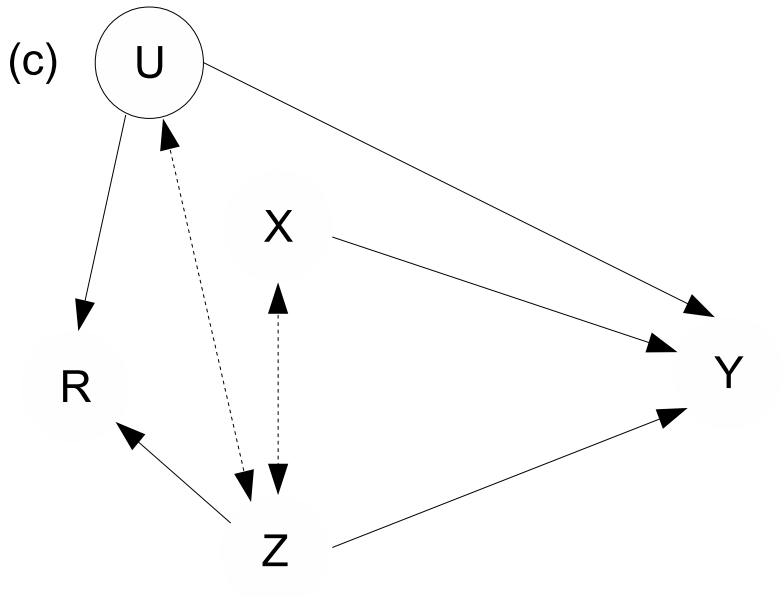
(a)



(b)

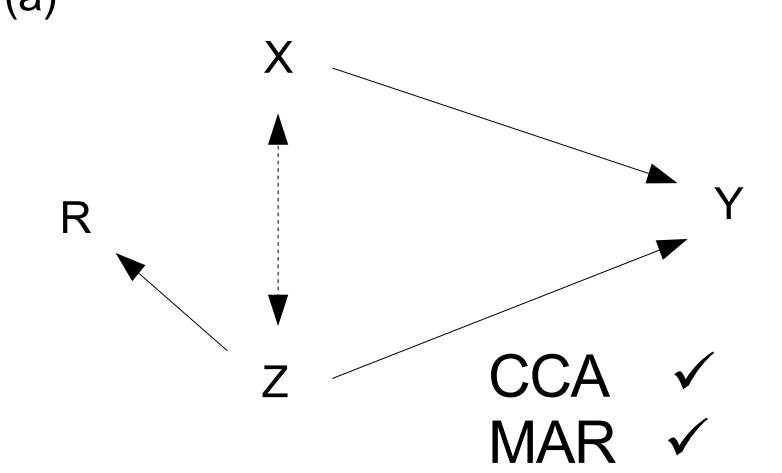


(c)

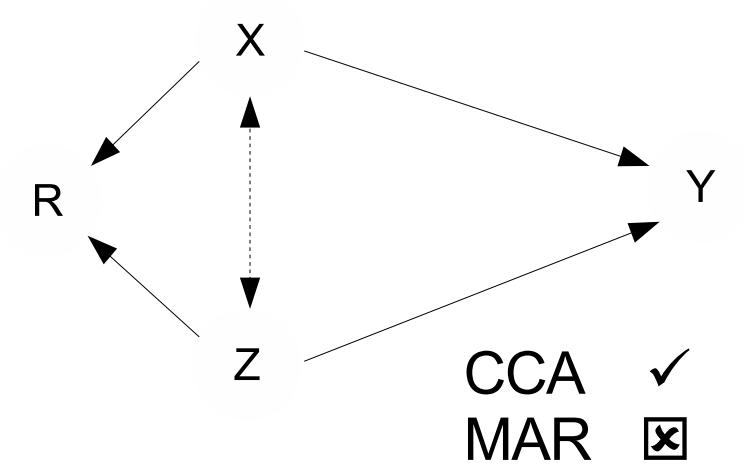


Covariate dependent missingness

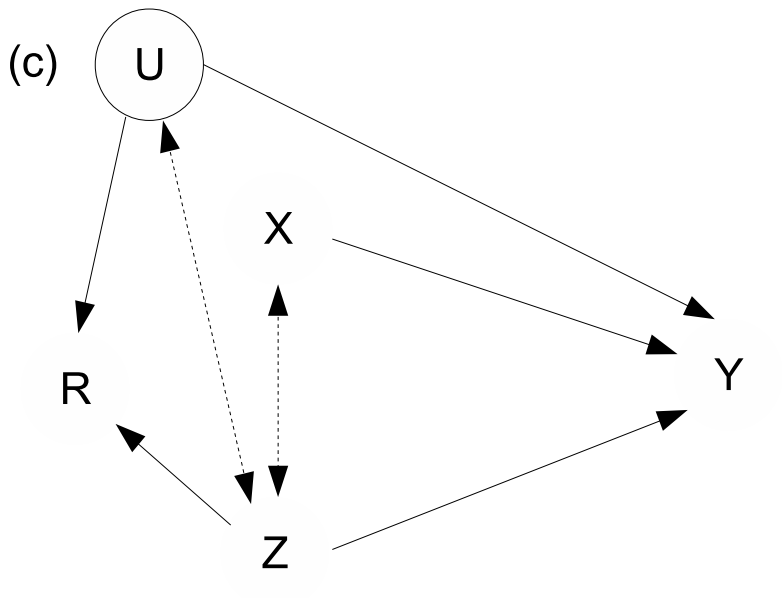
(a)



(b)

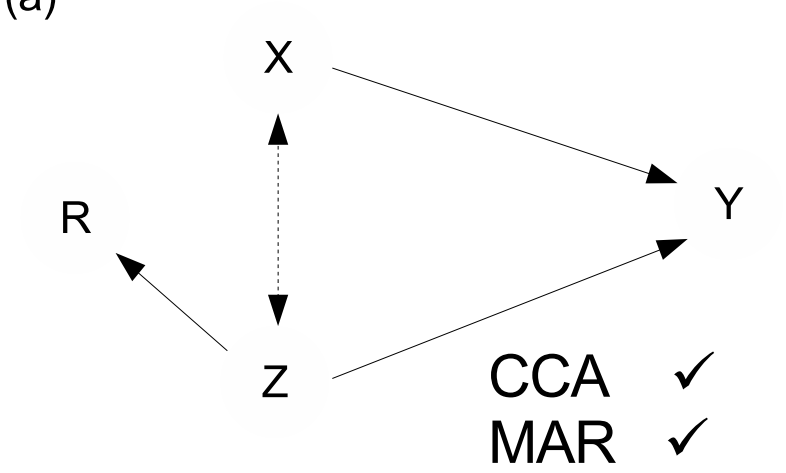


(c)

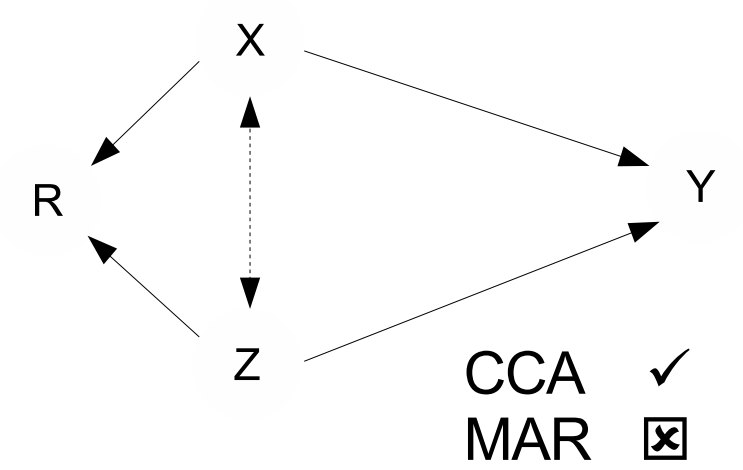


Covariate dependent missingness

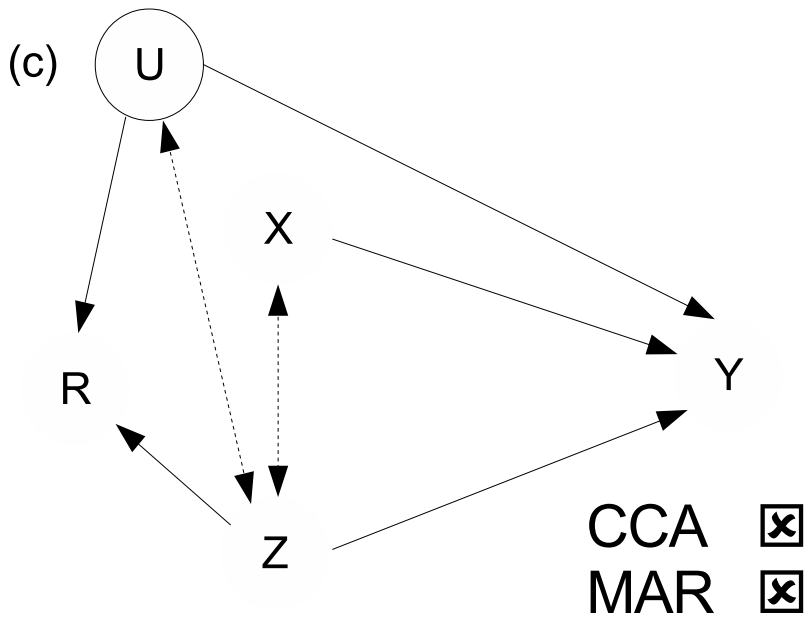
(a)



(b)



(c)



Covariate dependent MNAR missingness

- Sometimes missingness in covariates may be dependent on the value of the covariate itself, i.e. MNAR.
- e.g. at baseline we ask individuals how many units of alcohol they drink per week.
- Then it may be that missingness in the alcohol variable is dependent on how much they drink.
- If, given the alcohol variable and the other covariates, missingness is independent of outcome, CCA is ok.
- But an MAR analysis, such as using multiple imputation, will be biased.

Outline

Improving on the efficiency of CCA

- Suppose we believe a covariate is MNAR (ruling out MAR methods), but that the CCA assumption is plausible.
- We could use CCA, but it is inefficient.
- CCA makes no use of the observed information on (Y, Z) in the incomplete cases.
- We aim to construct an estimator which improves upon the efficiency of CCA.

Setup

- We assume interest lies in a parameter β indexing a conditional mean model for outcome Y given covariates X and Z:

$$E(Y|X,Z) = g(X,Z;\beta).$$

- Linear and logistic regression models are examples of conditional mean models (although they are fully parametric).
- We assume Y and Z are fully observed, but X is partially observed.
- We let R denote a binary indicator for whether X is observed ($R = 1$) or not ($R = 0$).

Estimation with full data

- In the absence of missing data, we can estimate β as the solution to estimating equations of the form [3]

$$\sum_{i=1}^{n} d(X_i, Z_i)\epsilon_i(\beta) = 0,$$

where $d(X_i, Z_i)$ is some function of (X_i, Z_i) and

$$\epsilon_i(\beta) = Y_i - g(X_i, Z_i; \beta).$$

- e.g. $d(X_i, Z_i) = (1, X_i, Z_i)^T$ results in the ordinary least squares estimator of β.
- The key to consistency is that $E(d(X_i, Z_i)\epsilon_i(\beta^*)) = 0$ where β^* denotes the true value of β.

Estimation with full data

- The estimating function has mean zero because

$$\begin{aligned} E(d(X,Z)\epsilon(\beta^*)) &= E(E(d(X,Z)\epsilon(\beta^*)|X,Z)) \\ &= E(d(X,Z)E(Y-g(X,Z;\beta^*)|X,Z)) \\ &= E(d(X,Z)\times 0) = 0. \end{aligned}$$

Complete case analysis estimating equation

- Complete case analysis consists of estimating β by solving the estimating equations

$$\sum_{i=1}^{n} R_i d(X_i, Z_i)\epsilon_i(\beta) = 0.$$

- The estimating function (and equation) continues to have mean zero if, as we assume, $R \perp\!\!\!\perp Y|X, Z$, since

$$\begin{aligned} &E(Rd(X,Z)\epsilon(\beta^*)) = E(E(Rd(X,Z)\epsilon(\beta^*)|X,Z)) \\ &= E(P(R=1|X,Z)d(X,Z)E(Y - g(X,Z;\beta^*)|X,Z)) = 0, \end{aligned}$$

since $E(Y|X,Z) = g(X,Z;\beta^*)$.

Improving on the efficiency of CCA

- To improve upon the efficiency of CCA we have to make additional assumptions.
- One approach involves specifying a model for the missingness mechanism, $P(R = 1|X, Z)$.
- Although estimation is possible, it is unattractive because the model for $R|X, Z$ cannot be fitted directly.
- It would thus be difficult to select an appropriate model for $R|X, Z$.

A model for $R|Y, Z$

- Instead, we will posit a model $P(R = 1|Y, Z; \alpha)$, with parameter α.
- Ordinarily we would use a logistic regression model, with outcome R, and covariates some function of Y and Z.
- **Note** this is not a model for the (assumed) true missingness mechanism, which is $P(R = 1|X, Z)$.
- Although we assume R and Y are independent given (X, Z), they are not independent conditional only on Z.
- This is because we pick up the effect of X on R indirectly via Y.

Augmented complete case estimating equation (ACC)

- Given our model $P(R = 1|Y, Z; \alpha)$, and an estimate $\hat{\alpha}$ we can estimate β as the solution to

$$\sum_{i=1}^{n} R_i d(X_i, Z_i)\epsilon_i(\beta) + (R_i - \hat{\pi}_i)\phi(Y_i, Z_i) = 0,$$

  where $\hat{\pi}_i = P(R_i = 1|Y_i, Z_i; \hat{\alpha})$.
- We have augmented the CCA estimating function by a term to which all subjects contribute.
- This augmentation term has mean zero provided the model $P(R = 1|Y, Z; \alpha)$ is correctly spcecified.

Efficiency – choosing $d(X, Z)$ and $\phi(Y, Z)$

- We gain efficiency because the incomplete cases now contribute to the estimation of β.
- The efficiency depends on our choices of the functions $d(X, Z)$ and $\phi(Y, Z)$.
- For simplicity, we choose $d(X, Z)$ as we would with full data. e.g. $d(X, Z) = (1, X, Z)^T$ for OLS.
- When α is known, the optimal choice is

$$\phi(Y, Z) = -E(d(X, Z)\epsilon(\beta^*)|R = 1, Y, Z).$$

Efficiency – choosing $d(X, Z)$ and $\phi(Y, Z)$

- When α has to be estimated, the optimal choice for $\phi(Y, Z)$ becomes more complicated.
- For simplicity, we use $\phi(Y, Z) = -E(d(X, Z)\epsilon | R = 1, Y, Z)$.
- This expectation depends on aspects of the joint distribution of the variables about which we have not made any assumptions, and β^*.
- We shall posit a model for $f(X|R = 1, Y, Z)$, and use this to approximate the above expectation.
- Important to note that if we get this model wrong, we will not introduce bias (although efficiency will be affected).

Algorithm for augmented complete case estimator

1. Posit model for $P(R|Y,Z)$, parametrized by α, and estimate via maximum likelihood, giving $\hat{\alpha}$.
2. Posit parametric model for $f(X|Y,Z)$, and fit this using complete cases ($R=1$)
3. Multiply impute X **for all subjects** m times, based on fitted model for $f(X|Y,Z)$. Let X_{ij} denote jth imputation of X_i.
4. Estimate β as the solution $\hat{\beta}$ to

$$\sum_{i=1}^{n} R_i d(X_i,Z_i)\epsilon_i(\hat{\beta}) + (R_i - \hat{\pi}_i)\hat{\phi}(Y_i,Z_i) = 0,$$

where

$$\hat{\phi}(Y_i,Z_i) = -\frac{1}{m}\sum_{j=1}^{m} d(X_{ij},Z_i)\epsilon_{ij}(\hat{\beta})$$

and

$$\epsilon_{ij}(\beta) = Y_i - g(X_{ij},Z_i,\beta).$$

Guaranteeing an efficiency improvement

- We actually use a slightly more complicated estimator.
- The modification we use ensures that we are guaranteed (asymptotically) to obtain an efficiency gain over CCA.

Outline

Simulation study

We simulated datasets for $n = 1,000$ independent subjects as follows:

$$\begin{aligned} \begin{pmatrix} X \\ Z \end{pmatrix} &\sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.25 \\ 0.25 & 1 \end{pmatrix} \right) \\ Y|X,Z &\sim N(\beta_0 + \beta_X X + \beta_Z Z, \sigma^2) \end{aligned}$$

with $\beta_0 = 0$, $\beta_X = 1$, $\beta_Z = 1$, and σ^2 chosen so that $R^2 = 0.1$.

Missingness mechanisms

Values of X were made missing according to two mechanisms:

- $P(R = 1|X, Z) = \text{expit}(X)$.
- $P(R = 1|X, Z) = \text{expit}(X - Z)$.

For both mechanisms $P(R = 1) = 0.5$ marginally.

Estimation methods

β_0, β_X and β_Z were estimated using:

- Complete case analysis.
- Multiple imputation ($m = 10$ imputations) assuming $X|Y, Z$ is normal and X is MAR.
- The proposed augmented complete case (ACC) approach, assuming $P(R|Y, Z) = \text{expit}(\alpha_0 + \alpha_Y Y + \alpha_Z Z)$, and using $m = 10$ imputations in Monte-Carlo integration.

Note that the assumed model for $P(R = 1|Y, Z)$ is (not exactly) correctly specified for the missingness mechanisms used.

Simulation results

Mean (SD) of estimates across 1,000 simulations

Estimator	$\beta_0 = 0$	$\beta_X = 1$	$\beta_Z = 1$
$P(R = 1\|X, Z) = \text{expit}(X)$			
CCA	-0.002 (0.243)	1.003 (0.237)	0.989 (0.231)
ACC	0.004 (0.244)	0.999 (0.234)	0.992 (0.169)
MI	-0.378 (0.182)	1.006 (0.238)	1.031 (0.166)
$P(R = 1\|X, Z) = \text{expit}(X - Z)$			
CCA	-0.008 (0.254)	1.007 (0.236)	0.987 (0.241)
ACC	-0.003 (0.251)	1.004 (0.232)	0.992 (0.197)
MI	-0.381 (0.184)	1.008 (0.237)	0.882 (0.183)

Simulation conclusions

- ACC approach is approximately unbiased, despite mild mis-specification of missingness model under assumed MNAR mechanism.
- At least for setup here, ACC improves efficiency for β_Z as much as MI.
- It does not recover as much information about β_0 as MI.
- Neither ACC nor MI recover information about β_X.
- With missingness dependent on X, MI is badly biased for β_0, little bias for β_X and small or moderate bias for β_Z.

Outline

NHANES example

- We illustrate the ACC approach with data from the 2003-2004 National Health and Nutrition Examination Survey (NHANES).
- We focus on a linear model for how SBP depends on reported average number of drinks consumed per day, adjusting for age and BMI.
- Of 2,111 participants, 720 (34.1%) are missing the alcohol variable.
- It seems a priori plausible that missingness in this is dependent on the number of drinks, and perhaps age (and BMI), and given these independent of SBP.
- In contrast, the MAR assumption is arguably implausible.

Model for $P(R = 1|Y, Z)$

Variable	Odds ratio (95% CI)	p-value
Age (decades above 50)	0.763 (0.723, 0.805)	< 0.001
BMI (kg/m^2)	0.978 (0.961, 0.996)	0.019
SBP (per 10mmHg above 125)	1.08 (1.01, 1.16)	0.020
SBP^2 (per 10 mmHg above 125)^2	0.979 (0.963, 0.996)	0.015

Those with low or high SBP have less chance of reporting alcohol consumption.

Consistent with heavy drinkers and tee-totalers being less likely to report alcohol consumption.

Estimation methods

- We estimated parameters of a linear model for SBP, with age, age^2, BMI and log(no. of drinks+1) as covariates.
- We used CCA, ACC, and MI assuming MAR.
- For ACC and MI, we used a negative binomial imputation model to impute missing no. of drinks, given the other variables.
- 200 imputations were used for both ACC and MI.
- Sandwich SEs for ACC, and Rubin's rules with sandwich variance estimator for MI.

Estimated dependence of SBP on age, BMI and no. of drinks

Estimates (SEs)

Variable	CCA	ACC	MI
Age (decades above 50)	3.94 (0.26)	3.87 (0.24)	3.88 (0.21)
Age^2 (decades above 50)^2	0.26 (0.14)	0.32 (0.11)	0.30 (0.12)
BMI (kg/m^2)	0.41 (0.080)	0.40 (0.065)	0.32 (0.070)
No. of drinks*	1.27 (0.58)	1.29 (0.58)	1.51 (0.65)
Constant**	-1.93 (0.80)	-2.13 (0.75)	-2.36 (0.81)

* $\log_e$(average no. drinks per day + 1)

** SBP centred at 125 mmHg

Conclusions

- Differences between MI and ACC are less dramatic here.
- Suggestion that MI is overestimating the alcohol effect, but difficult to be sure.
- Estimated intercept from MI is also somewhat larger.
- Estimates from ACC are more precise than from CCA.

Outline

Conclusions

- Covariate dependent missingness is sometimes more plausible than MAR.
- MI assuming MAR can be badly biased for the intercept parameter when a covariate is MNAR, independent of outcome.
- Our approach improves upon CCA efficiency.
- But one should be careful about specifying the model for $P(R = 1|Y, Z)$ correctly.
- Stata software for linear models: `net from http://missingdata.lshtm.ac.uk/stata` then select AUGCCA.
- Extensions to multivariate X, with non-monotone missingness, should be possible.

References I

[1] I R White and J B Carlin.

Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values.

Statistics in Medicine, 28:2920–2931, 2010.

[2] R M Daniel, M G Kenward, S N Cousens, and B L de Stavola.

Using causal diagrams to guide analysis in missing data problems.

Statistical Methods in Medical Research, 21:243–256, 2012.

[3] A Rotnitzky and J M Robins.

Analysis of semi-parametric regression models with nonignorable nonresponse.

Statistics in Medicine, 16:81–102, 1997.

[4] J W Bartlett, J R Carpenter, K Tilling, and S Vansteelandt.

Improving upon the efficiency of complete case analysis when covariates are MNAR.

under review, 2013.