Missing data

P.Romundstad

Eksempel: psykiatrisk RCT

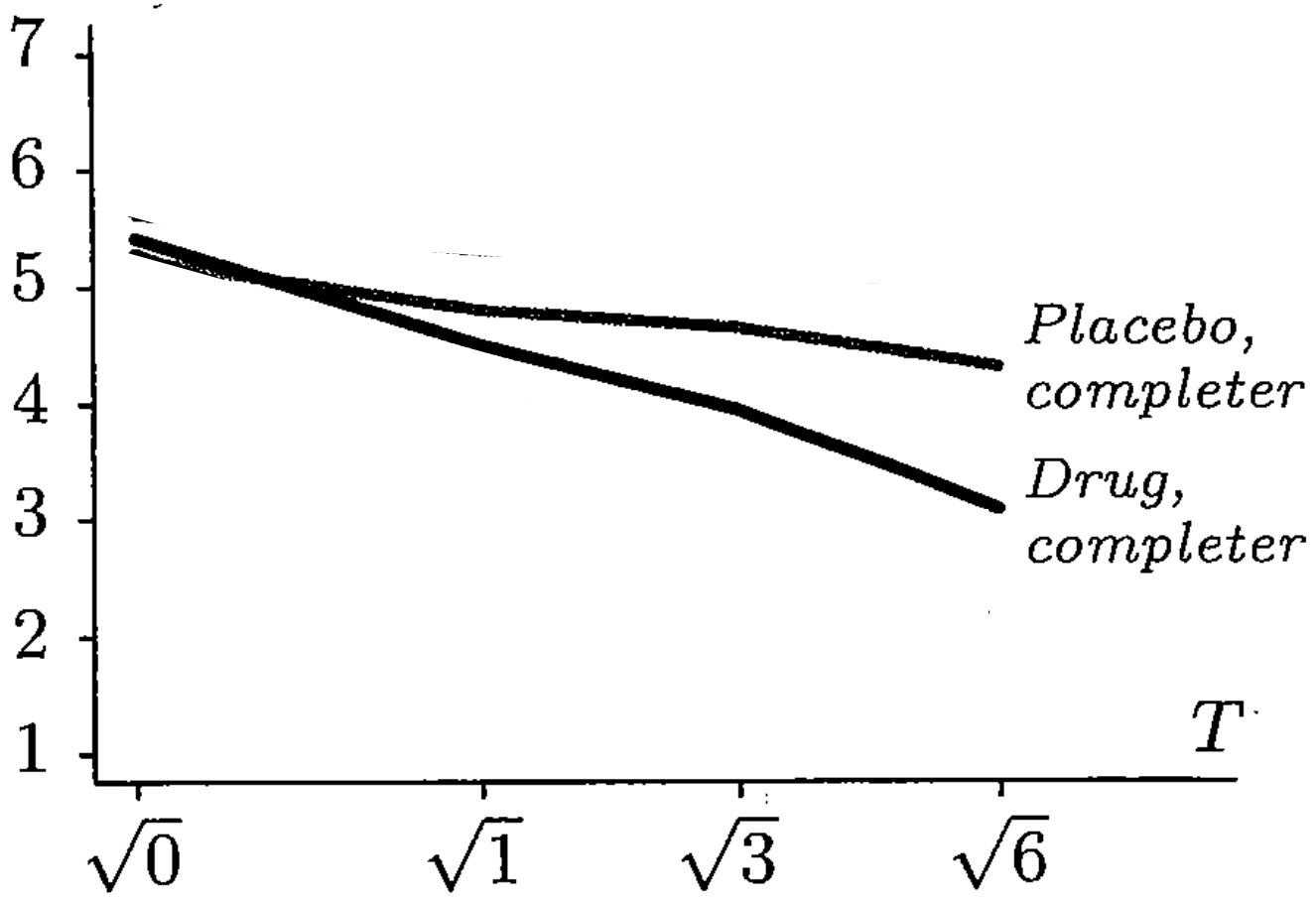
- Randomisert kontrollert trial:
- studere ny medisin for schizophreni
  - (Hedeker & Gibbons, Psychological methods 1997)

- 312 randomisert; 101 til placebo
- Målt sykelighet ved 0, 1, 3, 6 uker
- Utkom = alvorlighetskala for sykelighet
  (1 = normal, …, 7 = ekstremt syk)
  - behandlet som kontinuerlig variabel i analysen

- Missing data primært pga drop-out

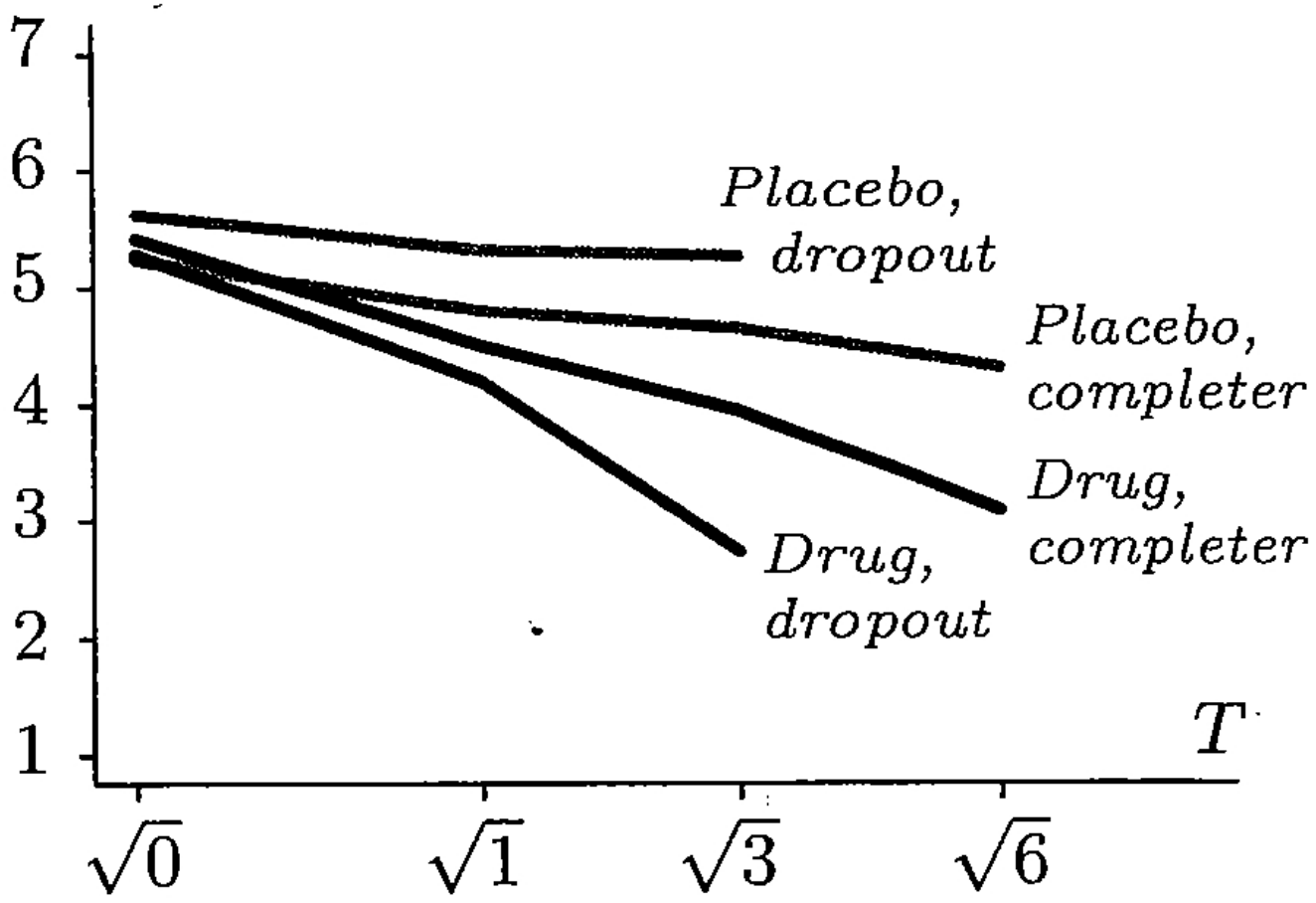
Eksempel: psykiatrisk RCT

- “Completers-only” analyse



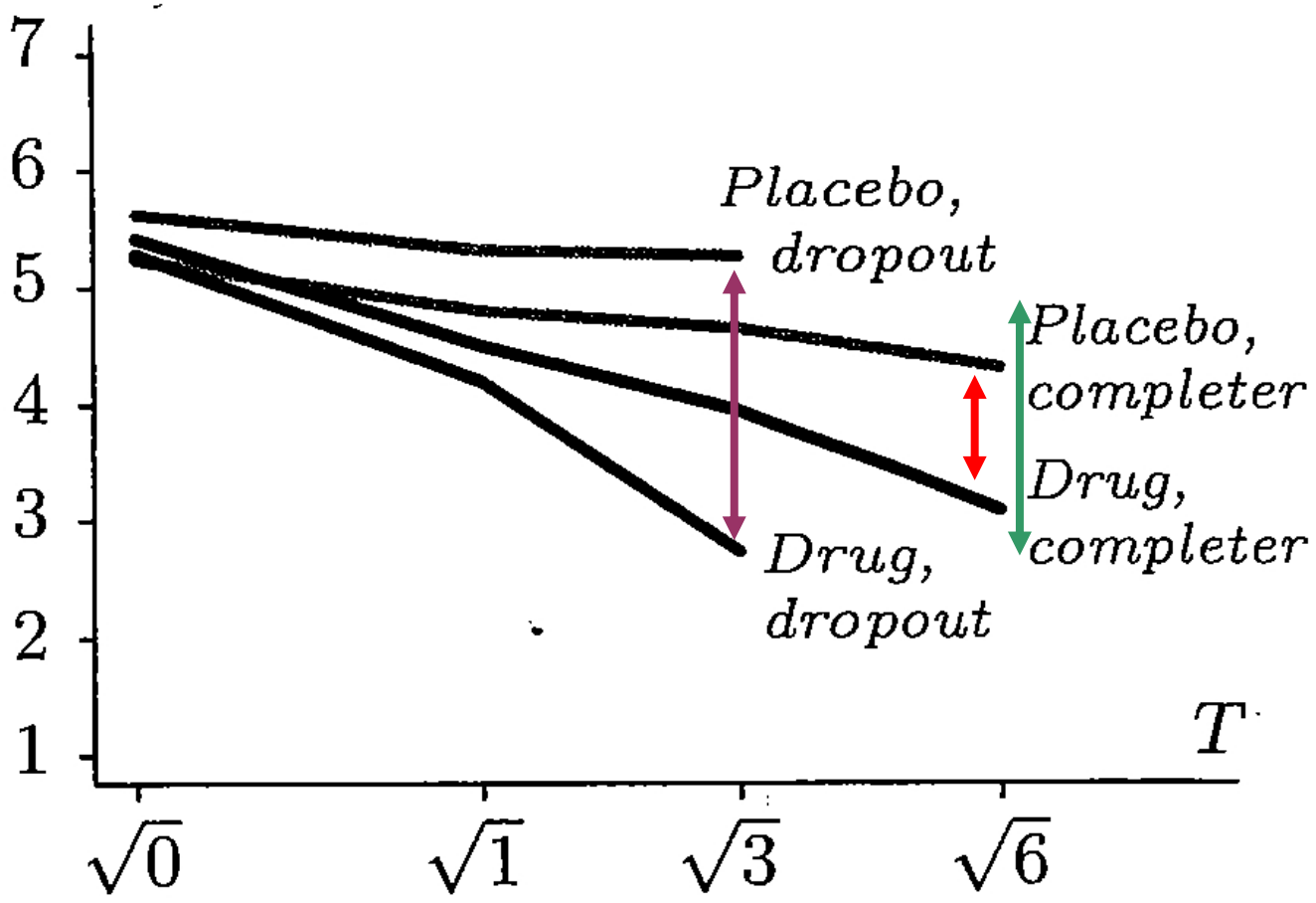
Eksempel: psykiatrisk RCT

- En “completers-only” analyse ville underslå behandelingseffekten betydelig:



Eksempel: psykiatrisk RCT

- En “completers-only” analyse wille underslå behandelingseffekten betydelig:



Konsekvenser av missing data

- De med missing kan være forskjellig fra respondenter (de uten missing data)

  → Kan gi BIAS ved analyse

- Missing data fører til mindre studiestørrelse (spesielt der en mangler data for flere faktorer (variabler)

  → Lavere presisjon

Årsaker til missing data

Wave non-response (går ut av studien underveis)

- Patient left the study because of illness
- Patient left the study because they felt better
- Participant emigrated

- Subject didn’t respond to one questionnaire because too depressed

Årsaker til missing

Item non-response

- Participant forgot to turn over the last page of the questionnaire so missed 5 questions
- Participant refused to disclose their income
- Participant didn’t respond to some sensitive questions on the depression scale
- Participant couldn’t remember their birth weight

Årsaker til missing

Generelt vil ikke årsaken være kjent med sikkerhet

Men, for å løse problemet med missing data må en forutsette visse underligende forklaringsmønstre:

1. MISSING COMPLETELY AT RANDOM (MCAR)
2. MISSING AT RANDOM -kovariat avhengig (MAR)
3. MISSING NOT AT RANDOM (MNAR)

Typer av missing

- ***Missing completely at random***

There are no systematic differences between the missing values and the observed values.
For example, blood pressure measurements may be missing because of breakdown of an automatic sphygmomanometer

- ***Missing at random***

Any systematic difference between the missing values and the observed values can be explained by differences in observed data.
For example, missing blood pressure measurements may be lower than measured blood pressures but only because younger people may be more likely to have missing blood pressure measurements

- ***Missing not at random***

Even after the observed data are taken into account, systematic Differences remain between the missing values and the observed values. *For example, people with high blood pressure may be more likely to miss clinic appointments because they have headaches*

Enkle løsninger kan være gale

- Complete case analyse
- Mean imputering (sette inn gjennomsnittet)
- Last value carried forward (LVCF)
- Missing kategori

- Alle disse metodene gir bias, utenom complete case analyse som ved ”missing completely at random” bare gir lavere presisjon

Hypotetisk eksempel

- Systolisk blodtrykk ved to tidspunkt (20år og 50 år)
- Hva er sammenhengen mellom blodtrykkene
- *Simulerte data-trekker tilfeldig fra en populasjon med følgende karakteristika (sanne parametre):*

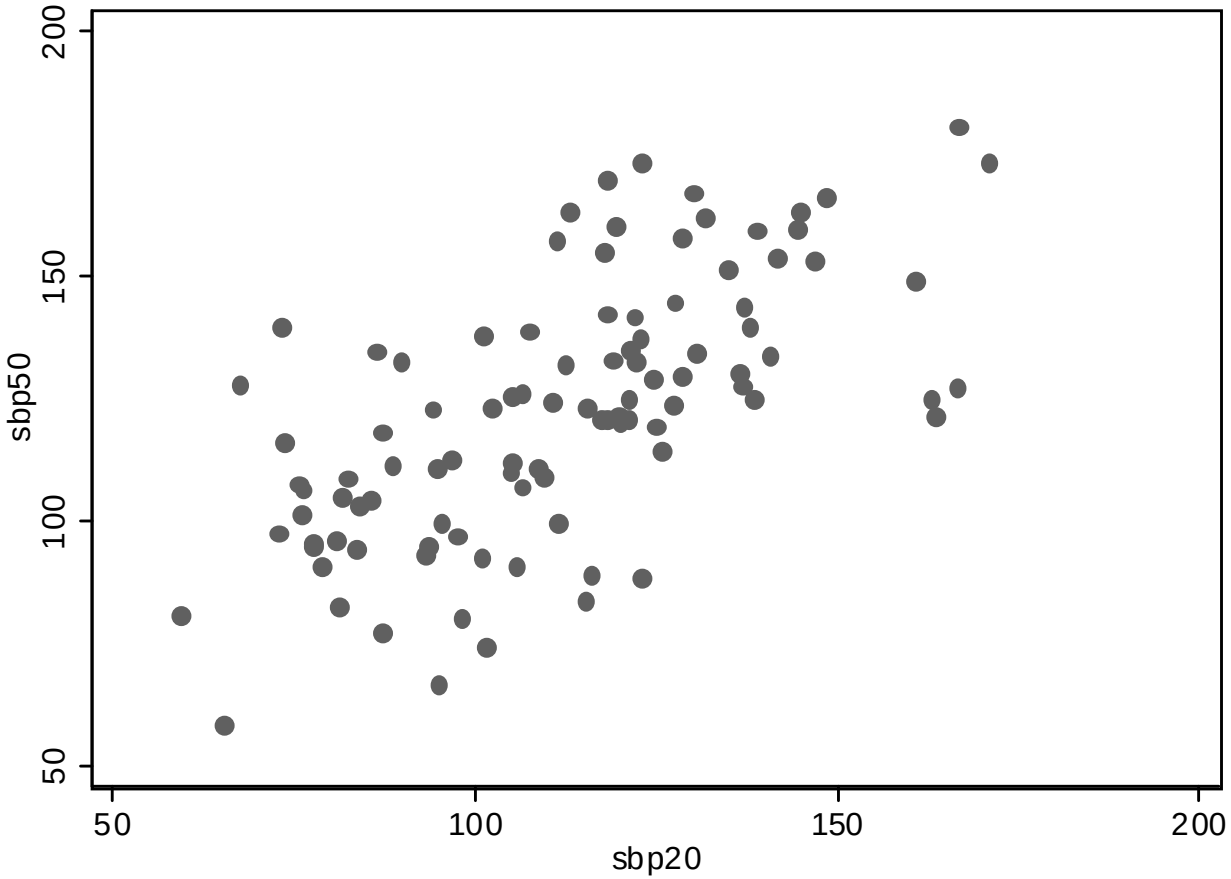
	X = SBP20	Y = SBP50
Mean	115	125
SD	25	25
Correlation = 0.6	$\beta_{Y\|X}$ =0.6	

One simulated dataset

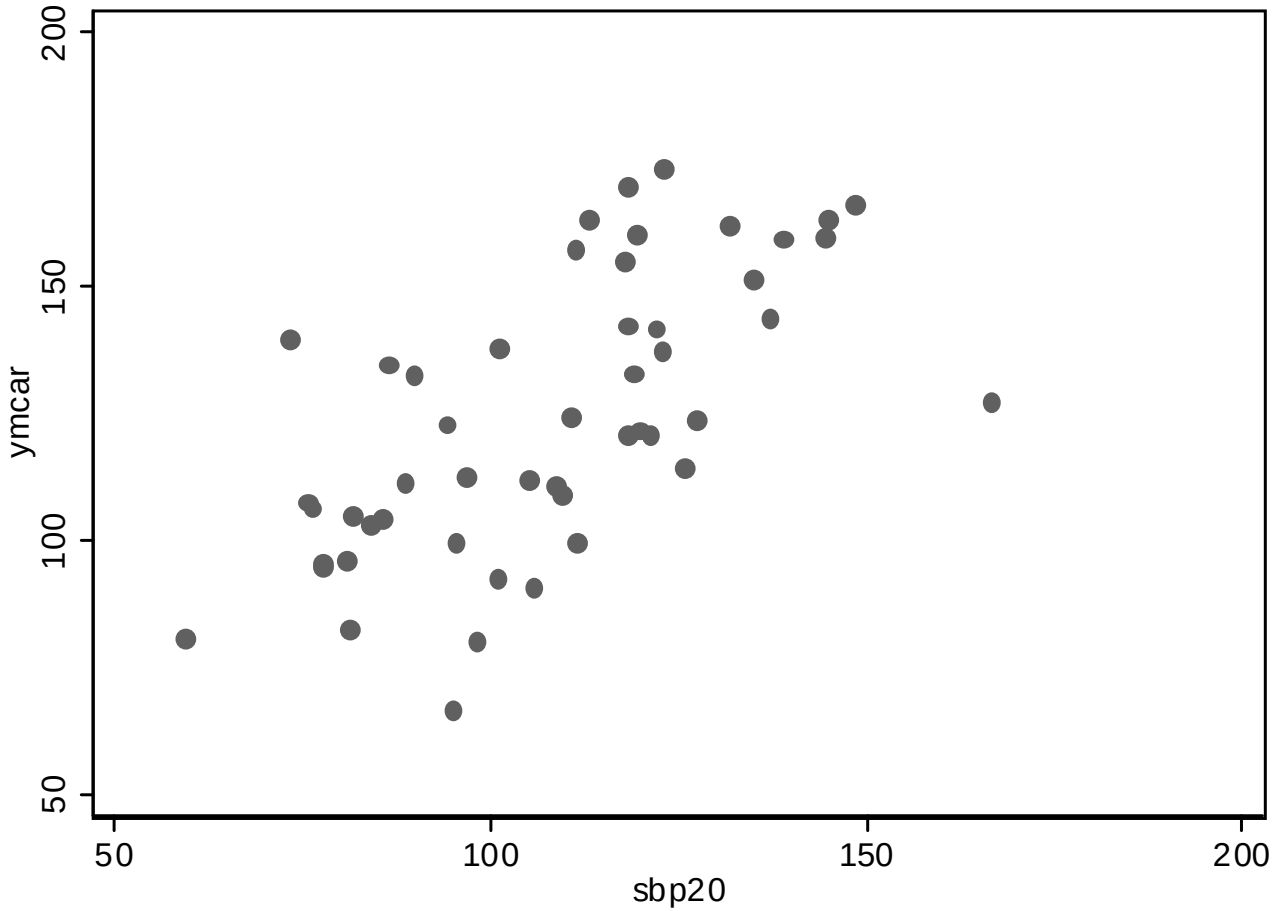
N=100

	SBP20	SBP50
Mean	111.6	122.5
SD	25.2	26.3

Corr = 0.66	$b_{Y\|X}$ = 0.67 (95%CI 0.51, 0.83)



- Lager missing data ved å tilfeldig fjerne 50% Y-verdier (i.e. SBP50) slik at vi får *Y_MCAR (“y missing compl. at random)*
- E.g. 50%: scatter plot ser omtrent likt ut som fullt data sett



Resultat med tilfeldig missing

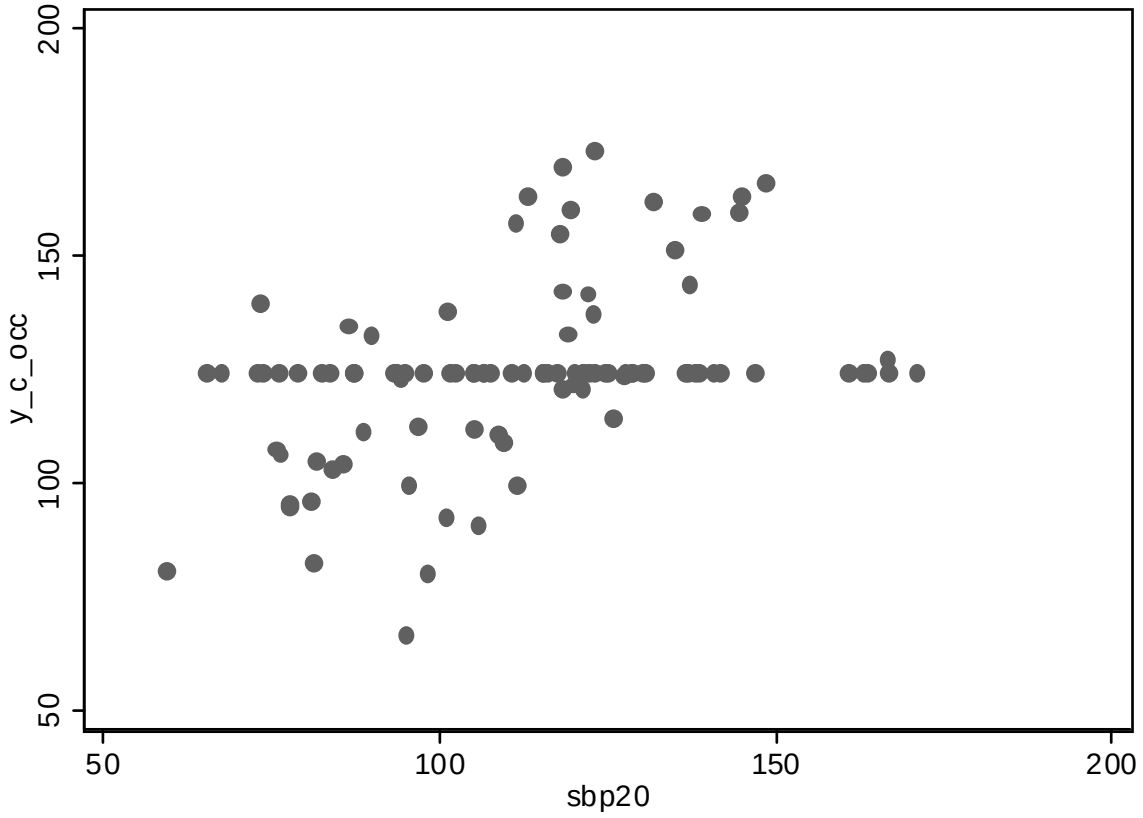
Data	Mean	SD	Corr	$b_{Y\|X}$	SE($b_{Y\|X}$)
Pop'n	125	25	0.6	0.6	-
Full sample	122.5	26.3	0.65	0.67	0.08
CC: 50% missing	124.3	27.3	0.65	0.77	0.13
CC: 78% missing	118.9	25.4	0.73	0.80	0.16

N.B. All estimates comfortably within 2 SE's of true beta; SEs ↑.

Mean imputation

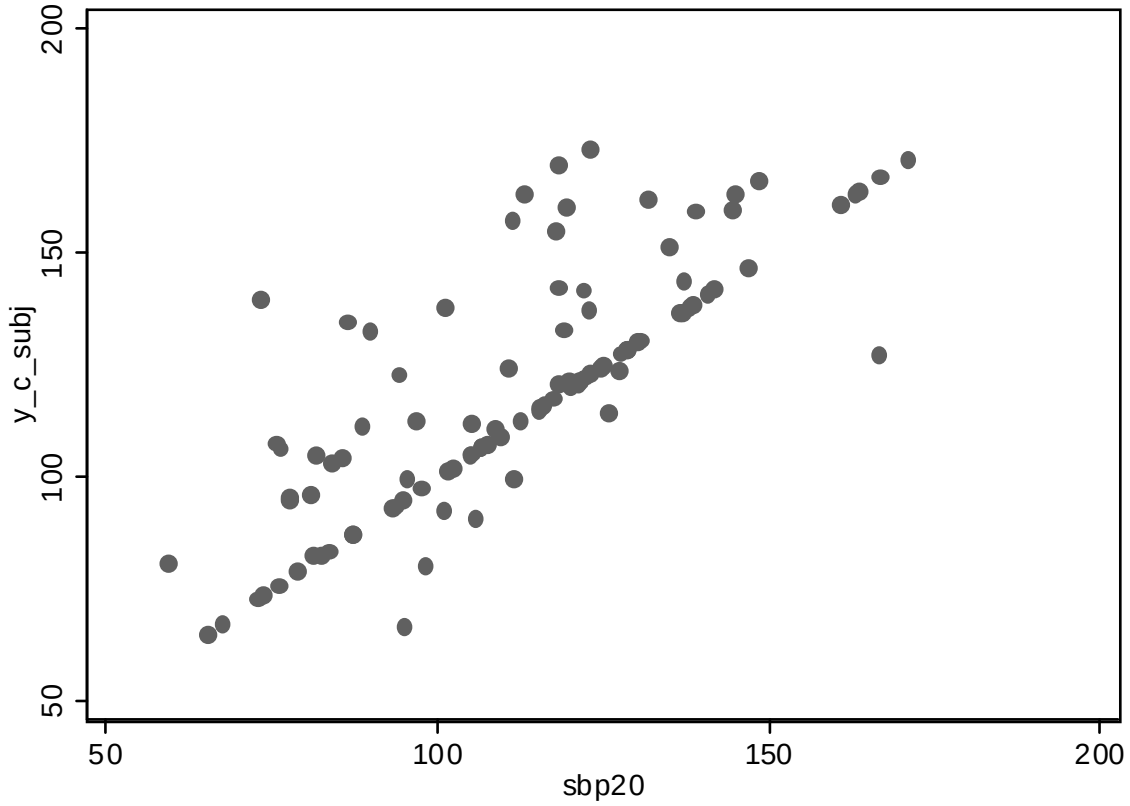
- Replace missing values with mean value of variable in question
- Several variations
  - Mean value all non-missing subjects
  - Subject mean value using other occasions of measurement
    - in this case, just first measure, i.e. *X* or SBP20 (LVCF)
  - Predicted mean from regression model of variable on other variable(s) (“conditional”)
    - here, regress *Y* on *X*, and substitute predicted *Y* where missing

- Example: imputing overall mean for missing *Y*



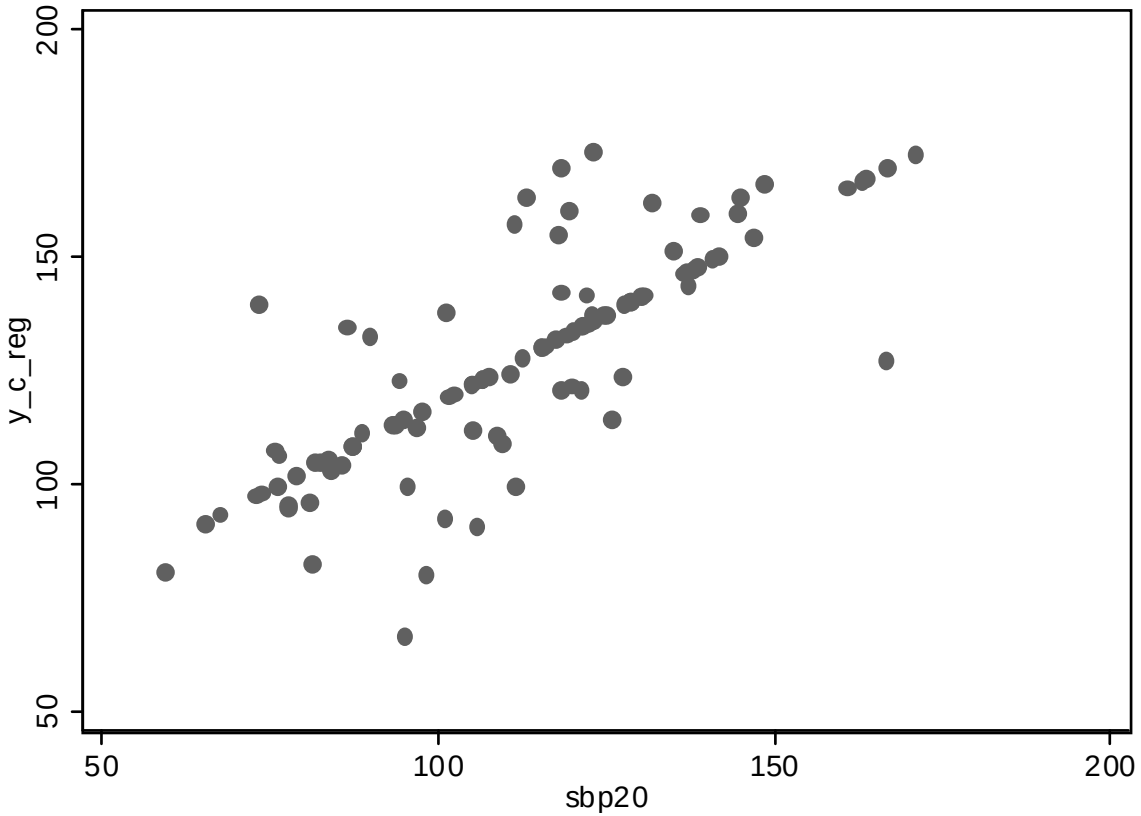
- Association attenuated

- Example: imputing baseline value (*X*) for missing *Y* *(in this case equivalent to LVCF)*



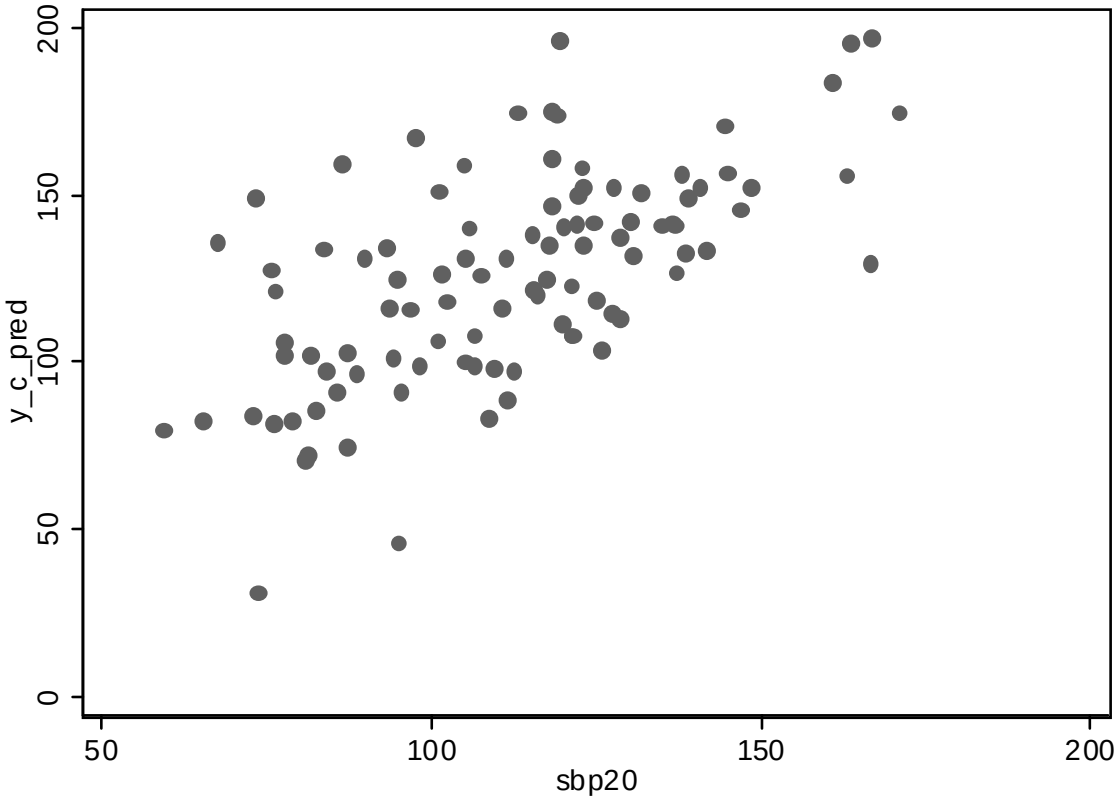
- Association biased (fails to allow for mean increase); uncertainty understated

- Example: imputing regression prediction $\hat{Y}$ for missing *Y*



- Association exaggerated; uncertainty understated

- Example: impute regression prediction "draw" $\hat{Y} + \text{error}$ for missing *Y*



- Recovers similar scatter to full data… but making up data at random?!

Data	Mean	SD	Corr	$b_{Y\|X}$	SE($b_{Y\|X}$)
Full sample	122.5	26.3	0.65	0.67	0.08
CC: 50% missing	124.3	27.3	0.65	0.77	0.13
Occasion mean	**124.3**	**19.2**	**0.42**	**0.32**	**0.07**
Subject mean	**119.9**	**27.3**	**0.79**	**0.85**	**0.07**
Regression prediction	**127.2**	**24.2**	**0.80**	**0.77**	**0.06**
Regression + error	**126.2**	**31.7**	**0.63**	**0.80**	**0.10**

Single imputing

- Imputing overall (“occasion”) mean distorts association and between-subject variation
- Imputing previous value (“subject” mean) seriously distorts trend over time (and within-subject corr)
- Imputing regression prediction preserves association but seriously understates variation
- Imputing regression prediction + error is unbiased but does not reflect true uncertainty (made-up data)
- **Conclusion**: single imputation is rarely useful and often dangerous

Last value carried forward

- Applied with repeated measures over time: impute values after drop-out with last observed value
  - Appealing when time-trends relatively stable
- Especially popular in clinical trials with longitudinal follow-up (e.g. FDA)
- Sometimes “conservative” in sense of understating differences in time-trends between groups
- However, also understates standard errors, so overall inference may not be conservative
- Especially bad for outcomes with high variation within subjects

Missing category indicator

- Popular method for handling missing values in *covariates* in a regression model
  (not used for outcomes with missing data)
- Create a dummy variable that indicates whether a covariate is missing
- Include this as an additional covariate
- Always biased!
- see Vach & Blattner *AJE* 1991

Muligheter ved missing data

- Hvor mange/andel missing?
- Undersøke årsak til missing
- Evaluere mulig betydning (tap av styrke - bias)

- Forutsette enten missing completely at random (MCAR), ”missing at random” (MAR) eller missing not at random (MNAR)

- Metoder for å ta hensyn til missing data ved MAR and MCAR
  - Complete case ved MCAR
  - Likelihood-based methods (ML), Eks: random effects for longitudinal data
  - Inverse-probability and doubly-robust weighting
  - Hot deck (Survey sampling)
  - Sensitivitetsanalyse
  - Multiple imputation

- Vanskelig ved MNAR

- Søk hjelp- dette kan gå galt!

What is multiple imputation (MI)?

Two stages:

a. Create *m* (≥ 2) *imputed* datasets with each missing value filled in
   - must be done “properly”, to incorporate relevant parameter uncertainty in imputed values

b. Analyse each imputed (complete) dataset using standard methods, and combine the results in appropriate way…

“Two-step” modelling

- Separation between imputation (to handle missing data) and analysis (to address original research questions)
  - Very appealing in practice, if enabling tools available
  - One-hit effort at modelling all variables subject to missingness (MAR assumption)
  - Model need not be entirely plausible for major gains to be made (Schafer)

“Two-step” modelling

- Allows data analyst to use standard complete-data analysis methods (with appropriate combination)
  - may even use non- or partially parametric methods
- One set of imputations may be used for many analyses
- Can be highly efficient even for small number of imputations

How to create MI datasets

- Draw from probability distribution of missing data given observed data

- Practical tools now available
  - Full model (e.g. multivariate normal)
    - Schafer's freeware (NORM etc)
    - SAS, S-PLUS
  - "Chained equations" ("sequential regressions")
    - MICE, IVEWARE, 'ICE' for Stata

How to analyse MI datasets

- Inference:
  Rubin's rules of combination…
  - Overall estimate = average of *m* separate estimates
  - Variance/SE: combines *within* and *between* imputation variance…

  "mim" package for Stata (Royston & Carlin)

MI principles:

- Key assumptions of “proper” imputation
  - Distribution of missingness: MAR
    (or need to propose explicit MNAR model)
  - Joint model to describe all data
  - Should preserve all aspects of data, in particular interaction effects
  - May use extra “auxiliary” variables
  - Prior distribution for parameters
- However
  - Model need only be approximately true…

Imputer's vs analyst's model

- In general, any variable(s) that may prove important in subsequent analysis should be present in imputation model
  - Includes using "future" to impute "past"

- ***Converse*** not at all necessary: If *Y* imputed under a model that includes *Z*, there is no need to include *Z* in future analyses involving *Y* unless the *Y*-*Z* relationship is of substantive interest.
  - Results concerning *Y* cannot be biased by inclusion of extra variables in the imputation phase.

- Rich imputation model to preserve maximum number of associations is desirable – may be used for wider variety of post-imputation analyses.

MI with auxiliary variables

- Suppose have variables of direct substantive interest $Y_1,\ldots,Y_p$ and another set $W_1,\ldots,W_q$ of "auxiliary" variables
- Should we include *W*'s in the imputation model?
  - When is it beneficial to include them?
  - When might it be harmful?
- Types of auxiliary variable

  A: correlated with outcome *Y* and with missingness

  B: correlated with *Y* only

  C: not correlated with either

MI with auxiliary variables

Conclusions from one simulation study
(Collins, Schafer & Kam, 2001)

- Omitting a correlate of missingness is not usually serious unless it is highly correlated with *Y* and there is a high rate of missingness
- Including correlates of outcome in model is **very** helpful; correlates of missingness rather less so
- Little danger in including too many variables: inclusive strategy better than restrictive
- These results favour MI over ML since more feasible to be inclusive with MI

Other imputation approaches

SAS Proc MI

- implementation of NORM

MLwiN macros: multilevel models

- day 3 if time permits (Carpenter et al, 2005)

Other imputation approaches

Nonparametric, semiparametric …

- Hotdeck
  - Related to traditional sample survey method for single imputation
  - Replace individual with missing values by a randomly drawn complete case, within strata (MCAR, MAR-CD)
  - Most useful with small amounts of missingness, in small number of variables
  - Stata add-in (Mander & Clayton)

Other major imputation approach: ICE

- Method attracting considerable attention currently
- Intuitively simple to understand, and flexible in implementation, but theoretical basis is controversial
- Original implementation in S-PLUS
  - Van Buuren (1999)
- Subsequent versions
  - SAS: Raghunathan et al (2001)
  - Now in Stata: Royston (2004, 2005), command named **`ice`**

ICE: outline of method

- ICE = Multiple Imputation using Chained Equations”

…also described as method of “regression switching”

1) Identify group of variables to be included
2) For each *variable* containing missing values, fit regression of *observed* values of *the variable with missing* on all other variables
3) Draw parameter values (β^*, σ^*) for this univariate regression model from approximate posterior distribution
4) Use these drawn parameter values to create new imputed value for missing cases

ICE: outline of method

5) Repeat this procedure for the next variable that contains missing values
6) Repeat the entire procedure for several cycles (van Buuren recommends 20, but fewer may suffice), to ensure convergence, and at the end save the currently imputed dataset
7) Repeat the whole process *m* times to obtain the required number of imputed datasets

ICE: further details

- Easy to implement since only requires univariate regression for each variable containing missing values

- Univariate regressions may be tailored appropriately, e.g.
  - logistic for binary outcomes
  - ordinal logistic for ordinal
  - multinomial logistic for categorical

ICE: further details

- Very flexible in representing associations in data
  - Can incorporate interaction terms in univariate regression specifications
  - Can include as many fully observed covariates as desired, with appropriate parameterisations
- Theoretical problem: how do we know it is valid??
  - Not based on an underlying joint model for the data
  - Possible that switching regressions may “converge” to mutually incompatible distributions
  - Flexibility may outweigh these concerns

ice in Stata

- Extensive default settings, to detect type of variable and determine appropriate regression accordingly (if binary, then logistic, etc)

- Provides full control for specification of univariate regressions
  - **eq** option: specify which variables to include of each univariate regression
  - o. and m. option provides easy implementation of categorical (multinomial and ordinal) variables