**Epidemiology: week 3**

**Practical:**

* MD\_practical\_II - questions

**Reading (same as week 2):**

* Groenwold et al 2011 Dealing with missing outcome data in RCT and observational studies
* Janssen 2010 Missing covariate data in medical research
* Liublinska 2011 Re -dealing with missing outcome data in RCTs and observational studies
* Naaktgeboren e.a. - 2016 - Anticipating missing reference standard data when planning diagnostic studies

**Summary:**

**Lecture 1: Imputation of Missing Data**

**Single (multivariable) regression based imputation**

Single imputation by regression for univariate missing data

* Includes outcome! -> preserves relation
* Includes all variables of the model
* Includes unknown predictors of the missing value
* Estimate actual value of a missing value given all other predictors

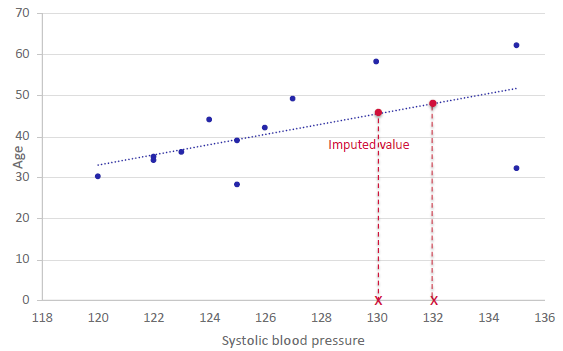
impmodel = glm(cvd ~ vacc + age + sex + pulm + DM + log(contact) + hosp,

family=binomial(), data=data)

then insert predicted value as imputation

data$cvd[is.na(data$cvd)] <- predict(impmodel,

newdata=data, type="response")[is.na(data$cvd)]

****

**Benefit:**

Uses all information (covariates and outcome)

**Problem:**

imputed values directly on regression line compared to observed values being scattered

**-> lacking variation**

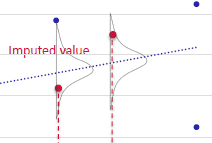
**Uncertainty of natural variation** **& uncertainty of estimated imputation model** (SE too low)

**Multivariate missing data** -> model cannot make use of observed data from other columns to calculate predicted values

**Solve problem by**:

**Adding natural variation** (binary case)

* Imputation is given by random sample from Bernoulli

**Adding natural variation** (continuous case)

* Adopt linear regression
* Yields residual variance
* Imputation by random sample from Normal
* adding variability by adding scatter based on the residual error -> imputed values NOT on regression line

**Standard Error is still underestimated (too significant)**

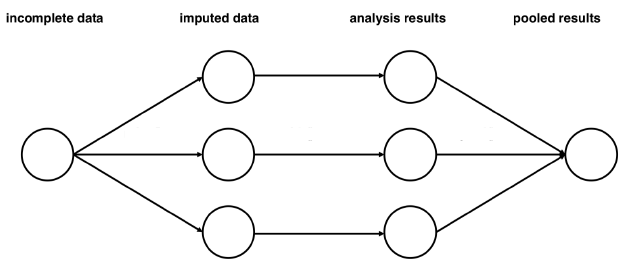
* we are **treating all data as if it was observed** data BUT the observed values are obviously more reliable than the imputed values and we are making an assumption by treating them the same

**Lecture 2: Imputation of Missing Data**

**Multiple imputation by regression**

Key change to single regression based imputation -> **multiple** imputations (duh)

* instead of stopping after one go we draw a random sample from estimated imputation model
* random sample of parameters(regression coefficients, residual error)
* **use sampled parameters to generate prediction**
* **add noise to prediction**
* repeat to generate multiple imputed datasets -> imputed values change over repetition



**Analysis and pooling of results**

* each computed dataset is now complete
* point estimate for each beta is average across m-sets -> unbiased under MAR and MCAR
* pooled standard error now has between imputation variation that introduces uncertainty to the data

**with multiple imputation regression:**

**MCAR**

* unbiased beta estimates (like SI)
* and unbiased standard error! As we accounted for uncertainty -> unbiased p-values

**MNAR**

* even NAs are partly MNAR imputation can reduce bias based on MAR

**Multivariate missingness**

* multiple imputation ineffective when cases with more than one NA
* because we cannot use observed values to estimate missing values or we would have to use estimate of values to estimate a value

Solution: iterative procedures

**MI by regression**

**recommendations:**

* should reflect all uncertainty (imputation model error & uncertainty on imputation model parameters)
* should be as flexible as possible (complexity for analysis model & interaction of model)
  + other variables may also carry information on NAs
* ALWAYS include the outcome (directly or indirectly)

**common pitfalls**:

* keeping output variable/ outcome out of the imputation
  + final analysis model is a model of the outcome predicted by several factors
  + **association between imputed predictor and outcome is lost when outcome is not included during the imputation**
    - leads to congeniality problem= the fact that all relations in analysis model should be represented in the imputation models
* non-normally distributed assumption (check setting of package)
* false assumption of MAR (e.g. falsely assuming that we have information to impute NAs)
* computational problems

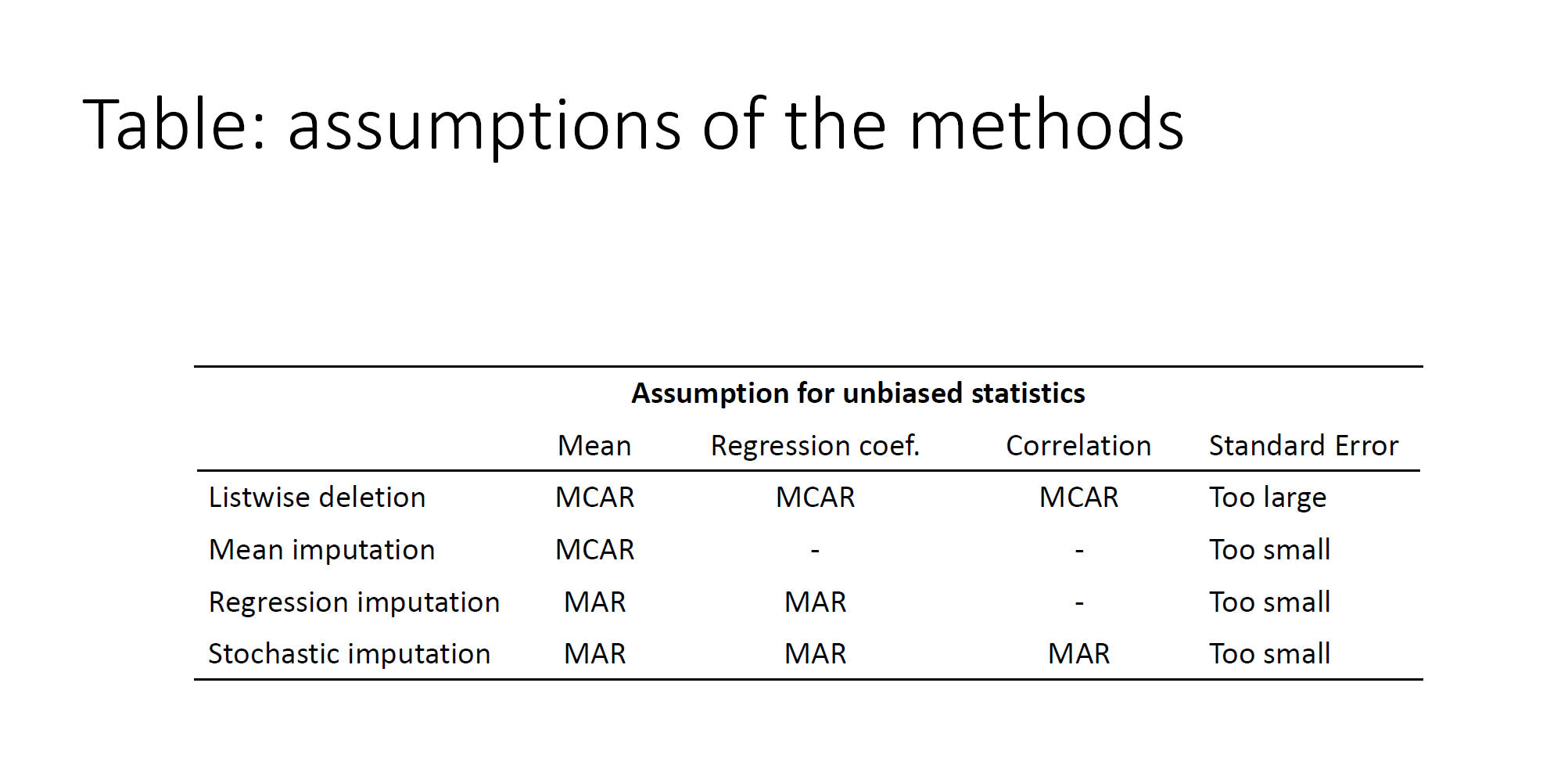
**Special case when only outcome data is missing & analysis based on max. likelihood:**

* unbiased & no imputation needed if:
  + CCA with adjusted covariate
  + Missing outcomes are MAR (NAs are conditional on data we have collected)
  + All covariates of outcome are included as covariates in the adjusted model

= fully adjusted model

**BUT**

* One can still reduce uncertainty by using imputation
* It allows for integration of post-randomization variable



**Q & A**

**Goals of imputation of missing data:**

* Maintain relationship in the data (correlation structure)
* Retain variation in the data
* Keep uncertainty in outcome that’s equal to the real uncertainty
* Keep Uncertainty in imputation model that’s equal to real uncertainty
  + Values of the coefficients for each covariance have to be reflected by imputation model
    - Draw values for each coefficient and generate imputations according to these values (beta)
* being able to impute missing values when NAs on multiple variables
  + input random value/mean and then iterate to improve the imputed values (there will always be uncertainty around the coefficient for each imputation)

**Goal is not to find best imputed values but to find the values that can be plotted best that maintain the variation and uncertainty of the given data**

**How many imputations should we do?**

M=50?! No clear answer.

With an infinite number of imputations we could catch all variation but as it’s not possible and imputations can be costly we have to take uncertainty into account to make up for smaller number of imputations -> more imputation = less correction later

**When should I use single imputation?**

Never 😊 The standard error is only smaller in single imputation because it is BIASED.