Why we shouldn't ignore measurement error and missingness in our data

Jonathan Bartlett

London School of Hygiene and Tropical Medicine

29th September 2010

Outline

Measurement error

Missing data

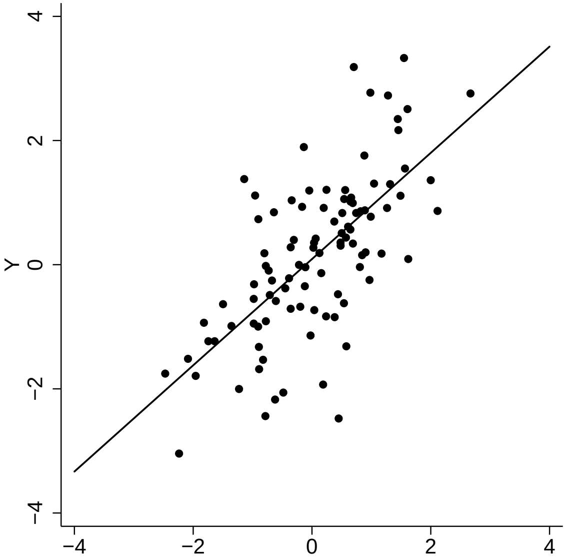
Measurement error

- For many quantities of interest in epidemiological and clinical studies, we may only be able to measure them imprecisely
- Such error has potentially important consequences for our subsequent analyses
- One example is measurement error in the (continuous) covariates of regression models

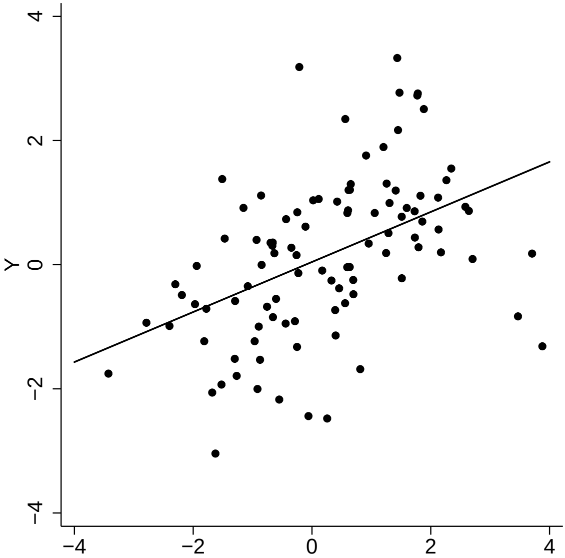
Measurement error in covariates

- With a single covariate, measurement error causes dilution in the estimated regression coefficient towards the null, and reduces our power to detect an association
- With multiple covariates, measurement error can bias estimates towards the null of away from the null
- If our confounders are measured with error, and we ignore this, our resulting estimates for the exposure of interest are not properly adjusted for the confounders

Attenuation due to covariate measurement error



Attenuation due to covariate measurement error



Example: the effect of salt intake on blood pressure

- Early studies investigating the association between salt intake and blood pressure resulted in somewhat contradictory conclusions
- Studies which used individual-level data often failed to find evidence of an association
- In contrast, population-level studies, in which average blood pressure in populations was related to average salt intake, did find evidence of an association

Example: the effect of salt intake on blood pressure

- This apparent lack of association from individual-level data was due to the fact that an estimate of an individual's salt intake based on a urine sample is a very noisy measurement of an individual's 'true' underlying salt intake
- Studies found that the ratio of between-subject variance to total variance (the reliability) for urine derived estimates of salt intake was around 0.33
- This means that if the true slope of blood pressure against salt intake is equal to β, estimates using individual-level data are diluted by a factor of 3

Example: the effect of salt intake on blood pressure

- After making a correction for the impact of measurement error in urine estimates of salt intake, estimates from individual-level data were in close agreement with estimates of β from population-level studies
- The apparent contradictory evidence was thus reconciled, once the effects of measurement error in urine sample estimates of salt were allowed for

Other examples

- Other studies have made similar corrections for the effects of blood pressure on CVD, which is also measured with considerable imprecision
- The result: the relative risk for the effect of blood pressure on CVD is much larger than was originally estimated

Correction for covariate measurement error

- We can correct for the bias induced by measurement error if we are able to estimate the magnitude of measurement errors
- To do this, we can take replicate measurements on a subset of our study subjects
- There are many different approaches to making the adjustment, but in many cases an adjustment can be made with hand-calculations or using appropriate programs in standard statistical packages (e.g. the cme command in Stata)

Outline

Measurement error

Missing data

Missing data

- Epidemiological and clinical studies invariably suffer from missing data, to a lesser or greater extent
- Missing data always result in a loss of information
- Perhaps more critically, missing data may introduce bias into our estimates

Missing data

- In the presence of missing data, we must make additional assumptions (implicit or explicit), in order to get valid estimates and inferences
- Large body of statistical methodology has been developed for handling missing data, e.g. multiple imputation
- The validity of the resulting estimates rest on some assumptions about the missing data
- Clever methods are not a panacea...

The QRISK study

- The QRISK study aimed to derive a new cardiovascular disease (CVD) risk score for the UK, based on routinely collected data from general practice
- The score was derived using data from 1.28 million patients registered at UK GP practices between 1995 and 2007, who were free from CVD at registration
- The outcome of interest was time to first recorded diagnosis of CVD
- Cox proportional hazards models were used to model time to CVD, as a function of risk factors measured at registration

Missing data in QRISK

- Inevitably there was substantial missingness in 'baseline' risk factor data
- In particular, 70% of subjects had HDL cholesterol missing
- A complete case analysis may be biased, if the complete cases are not representative of overall population of interest
- Complete case analysis are also inefficient
- The investigators therefore used multiple imputation to deal with missing data, using the ICE command in Stata

Cholesterol and CVD

- In the final model, the adjusted hazard ratio for the ratio of total to HDL cholesterol was 1.001 (95% 0.999 to 1.002)
- The apparent lack of an association between cholesterol and CVD was unexpected

Cholesterol and CVD

- A complete case analysis did show evidence for an effect of cholesterol
- It turned out that when imputing the missing values, although the time to CVD or censoring was included in the imputation model, the event indicator (1=CVD, 0=censored) had inadvertently not been used
- The imputed cholesterol values thus did not have the correct association with time to CVD, resulting in there being no evidence of an effect
- Re-running with a more appropriate imputation model, an association with cholesterol was found

Handling missing data

- Analyses in the presence of missing data introduce additional ambiguity into the analysis
- Methods such as multiple imputation can often help, by providing estimates which are valid under weaker assumptions than complete case analysis
- However, such methods should not be treated black-box algorithms – to get reasonable estimates out we must think carefully about the models and assumptions which are being used / made