Correcting for covariate measurement error using regression calibration

Centre for Statistical Methodology Forum

Jonathan Bartlett

London School of Hygiene and Tropical Medicine

14th October 2011

Outline

The effects of covariate measurement error

Regression calibration

Our model of interest

- Suppose we are interested in estimating the relationship between an outcome Y and covariates/explanatory variables $X_1, .., X_p$.
- We might do this by fitting a regression model (e.g. linear regression, logistic regression, Cox proportional hazards model).
- e.g. linear regression:

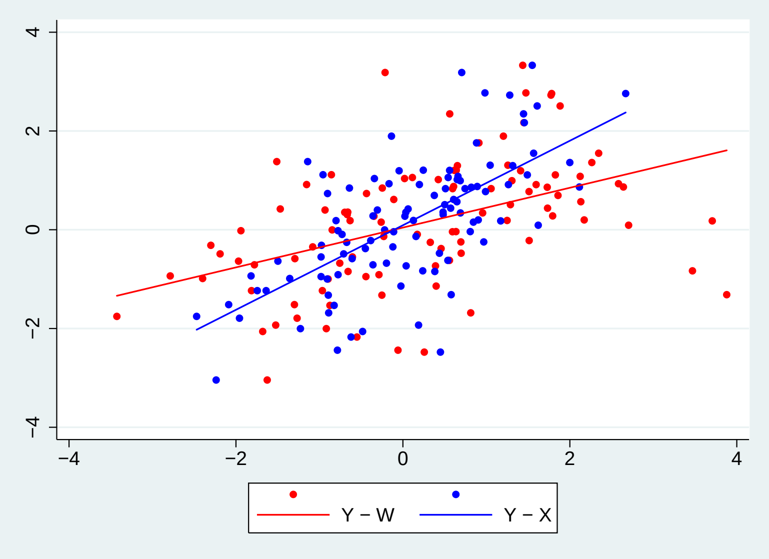
$$Y = \alpha + \beta_1 X_1 + ... + \beta_p X_p + \epsilon$$

where $\epsilon \overset{\text{iid}}{\sim} N(0, \sigma_\epsilon^2)$.

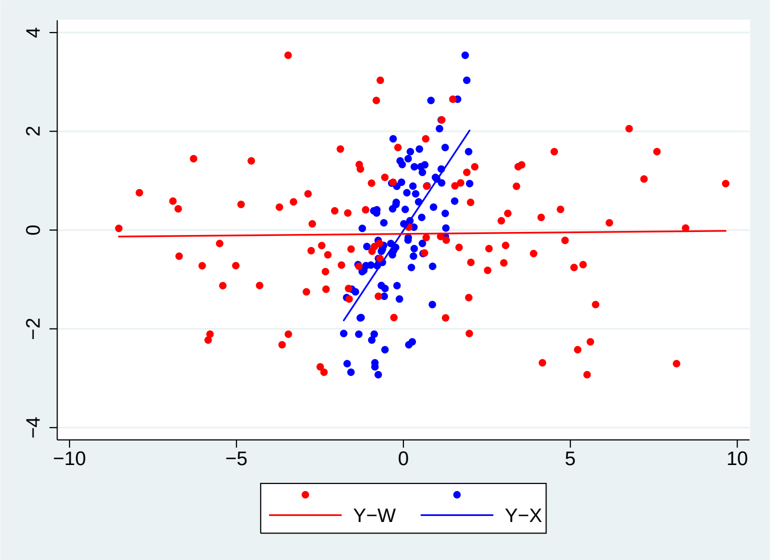
Covariate measurement error

- For some explanatory variables, we may only have a noisy error-prone measurement W of the true exposure/confounder of interest.
- e.g. we have single measurements of blood pressure for each subject, but we are interested in $X =$subject's 'true' blood pressure at entry to our study.
- What effects does such noise/error have?
- One might think that completely random noise may reduce precision, but not bias our estimates.

The effects of covariate measurement error in linear regression



More extreme measurement error



The effects of covariate measurement error in linear regression

- We will continue in this simplified setting:

$$Y = \alpha + \beta X + \epsilon$$

- Suppose that we actually measure $W = X + U$, where U is an independent measurement error.
- This is called the classical measurement error model.
- What slope β^* do we estimate if we fit the model with W as the covariate, instead of X?
- One can show:

$$\beta^* = \beta \frac{Var(X)}{Var(X) + Var(U)}$$

The effects of covariate measurement error in linear regression

$$\beta^* = \beta \frac{Var(X)}{Var(X) + Var(U)}$$

- The fraction on the right is what is referred to as the reliability of the measurements.
- As the amount of noise/error ($Var(U)$) increases, the slope is attenuated by a larger amount.
- The residual error in the regression also gets larger, so we have reduced power to detect associations.
- With multiple covariates, biases can be both towards **no** association (attenuation) or towards **larger** association.

Implications for epidemiology

- In the absence of confounders, exposure to disease associations are underestimated in magnitude.
- When we have confounders in our model, some of which are measured with error, our exposure to disease associations will **not** be properly adjusted for the confounders.
- When trying to determine which variables are most important in determining an outcome, variables can seem less important simply because they are measured imprecisely / with lots of error.

Allowing for covariate measurement error

- There are a number of statistical approaches to allowing for the effects of covariate measurement error.
- In addition to some assumptions about the nature of the errors (e.g. independence), all of them require some information on the magnitude of the errors.
- To quantify the magnitude of errors, we either need:
  - (Some) subjects to have repeated error-prone measurements, e.g. W_1 and W_2, or
  - (Some) subjects to have the true X measured, in addition to W.
- We will focus on the first situation.
- Such data allow us to estimate the magnitude of errors ($Var(U)$) and the variability of the true covariate ($Var(X)$).

Outline

The effects of covariate measurement error

Regression calibration

Regression calibration

- Regression calibration (RC) is one statistical approach for correcting for the effects of covariate measurement error.
- Rather than using W as the covariate in our model, we use $E(X|W)$ as covariate.
- $E(X|W)$ is the predicted value of X based on the error-prone measurement W.
- For classical error, $E(X|W) \neq W$!

Why RC works

- As before, suppose:

$$Y = \alpha + \beta X + \epsilon$$

- Then taking expectations conditional on W:

$$E(Y|W) = \alpha + \beta E(X|W),$$

  provided ϵ is independent of W (the non-differential error assumption).
- This result means that if we regress Y on with $E(X|W)$ as covariate, we obtain unbiased estimates of β.

Implementing RC

Implementing RC consists of the following steps:

1. assume a model for X and W,
2. fit this model using data (usually from the study in question) where some subjects have repeat error-prone measurements,
3. calculate $\widehat{E}(X|W)$ based on the estimated parameters,
4. run the regression model using $\widehat{E}(X|W)$ as the covariate,
5. adjust standard errors & confidence intervals for estimation of $\widehat{E}(X|W)$.

Fitting a model for repeated error-prone measurements

- Suppose we have data where subjects have $W_1 = X + U_1$, and some subjects have a second measurement $W_2 = X + U_2$.
- We assume $X \sim N(\mu_X, Var(X))$ and that $U_1, U_2 \stackrel{\text{iid}}{\sim} N(0, Var(U))$.
- This is the standard one-way analysis of variance or random-intercepts model.
- Fitting this model (e.g. using `loneway` or `xtmixed` in Stata) gives us estimates of μ_X, $Var(X)$, and $Var(U)$.

Calculating $\widehat{E}(X|W)$

- Under this assumed model:

$$E(X|W) = \mu_X + \frac{Var(X)}{Var(X) + Var(U)}(W - \mu_X)$$

- For those subjects with multiple error-prone measurements, we can use the mean of their error-prone measurements $\overline{W}$ to predict X:

$$E(X|\overline{W}) = \mu_X + \frac{Var(X)}{Var(X) + Var(U)/k}(\overline{W} - \mu_X)$$

when $\overline{W}$ is the mean of k measurements.

Stata commands for this simple setting

```
reshape long w, i(id) j(obs)
xtmixed w || id:
predict x_pred, fitted
reshape wide
reg y x_pred
```

Inference and other settings

- The method generalises to the setting of multiple explanatory variables measured with error, and to when we have some explanatory variables measured without error (e.g. gender, treatment group).
- Standard errors and confidence intervals should allow for first stage of estimation, e.g. by using bootstrap methods.
- RC also gives approximately unbiased estimates for some other model types (logistic regression, Cox regression).

References

[1] R. J. Carroll, D. Ruppert, L. A. Stefanski, and C. M. Crainiceanu.
Measurement Error in Nonlinear Models.
Chapman & Hall/CRC, 2nd edition, 2006.

[2] J. W. Hardin, H. Schmiediche, and R. J. Carroll.
The regression calibration method for fitting generalized linear models with additive measurement error.
Stata Journal, 3:361–372, 2003.