Linear Regression Models
P8111

Lecture 24

HW6

5/3

FP

4/28

5/12

Jeff Goldsmith

April 21, 2016

THE DEPARTMENT OF BIOSTATISTICS
Columbia University
MAILMAN SCHOOL
OF PUBLIC HEALTH

Today's Lecture

$E(y_i \mid x_i)$

$\underline{y_i} = \beta_0 + \beta_1 x_i + \epsilon_i$

- Measurement error in predictors
  - Impact
  - Approaches
- Mediation and confounding

Simple linear regression

We started with the model

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

where

$$\epsilon_i \sim \mathrm{N}\left[0, \sigma_\epsilon^2\right]$$

Throughout, we have been concerned with variability on the y_i.

- Biological variability
- Measurement error

Simple linear regression

Sometimes, the x_i are also observed with error

$$w_i = x_i + u_i$$

where x_i and u_i (and ϵ_i) are independent, and

$$u_i \sim \mathrm{N}\left[0, \sigma_u^2\right]$$

- May also be measurement error
- Surrogate variable error – using one variable for all subjects in a region
- Error induced by definition – yesterday's caloric intake to represent exposure

Full model

Classical measurement error model

$$\begin{aligned} y_i|x_i &= \beta_0 + \beta_1 x_i + \epsilon_i \\ \boxed{w_i = x_i + u_i} & \\ u_i &\sim \mathrm{N}\left[0, \sigma_u^2\right] \\ \epsilon_i &\sim \mathrm{N}\left[0, \sigma_\epsilon^2\right] \end{aligned}$$

with (x_i, u_i, ϵ_i) all independent

Identifiability issues

$x_1 \ldots x_n$

$w_1 \ldots w_n$

$u_1 \ldots u_n$

For the full model,

$$(\beta_0, \beta_1, \mu_x, \sigma_x^2, \sigma_u^2, \sigma_\epsilon^2)$$

and

$$(\beta_0, \beta_1, \mu_x, \sigma_x^2 + \sigma_u^2, 0, \sigma_\epsilon^2)$$

yield identical distributions, i.e. the model is not identifiable.
We need more information

- σ_u^2 (or, at least, $\hat{\sigma}_u^2$)

Regression

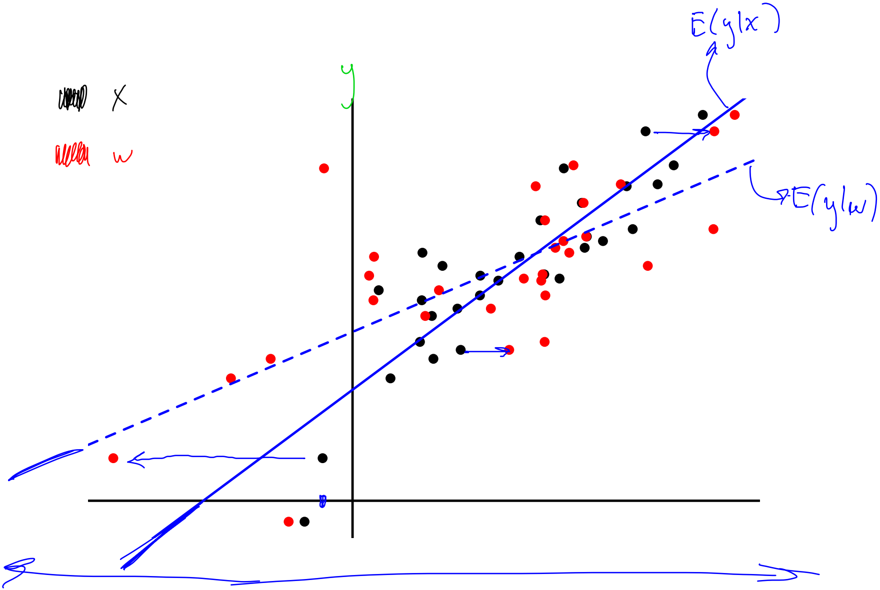
Still want to know

$$E(y_i|x_i) = \beta_0 + \beta_1 x_i$$

- We observe w_i rather than x_i
- What if we just use OLS?

$$E(y_i|w_i) = \beta_0^* + \beta_1^* w_i$$

OLS



Observed regression

$$\frac{\Sigma\,(x\ldots\ldots)(y\ \ldots\ldots)}{\Sigma\,(x\ldots\ldots)^2}$$

$$\begin{aligned}
\boxed{\hat{\beta}_1^*} &= \frac{\hat{\sigma}_{y,w}}{\hat{\sigma}_w^2} \\
&= \frac{\hat{\sigma}_{y,x}}{\hat{\sigma}_x^2 + \hat{\sigma}_u^2} \\
&= \boxed{\frac{\hat{\sigma}_{y,x}}{\hat{\sigma}_x^2}} \frac{\hat{\sigma}_x^2}{\hat{\sigma}_x^2 + \hat{\sigma}_u^2} \\
&= \hat{\beta}_1 \boxed{\frac{\hat{\sigma}_x^2}{\hat{\sigma}_x^2 + \hat{\sigma}_u^2}}
\end{aligned}$$

Observed regression

$$
\begin{aligned}
\hat{\beta}_0^* &= \hat{\mu}_y - \hat{\beta}_1^* \hat{\mu}_w \\
&= (\hat{\beta}_0 + \hat{\beta}_1 \hat{\mu}_x) - \left(\hat{\beta}_1 \frac{\hat{\sigma}_x^2}{\hat{\sigma}_x^2 + \hat{\sigma}_u^2} \right) \hat{\mu}_x \\
&= \hat{\beta}_0 + \hat{\beta}_1 \left(1 - \frac{\hat{\sigma}_x^2}{\hat{\sigma}_x^2 + \hat{\sigma}_u^2} \right) \hat{\mu}_x
\end{aligned}
$$

Attenuation correction

$$\hat{\beta}_1^* = \hat{\beta}_1 \frac{\hat{\sigma}_x^2}{\hat{\sigma}_x^2 + \hat{\sigma}_u^2}$$

We can use

$$\begin{aligned}
\boxed{\hat{\beta}_1} &= \hat{\beta}_1^* \frac{\hat{\sigma}_x^2 + \hat{\sigma}_u^2}{\hat{\sigma}_x^2} \\
&= \hat{\beta}_1^* \frac{\hat{\sigma}_w^2}{\hat{\sigma}_w^2 - \boxed{\hat{\sigma}_u^2}}
\end{aligned}$$

Regression calibration

$$X_i = w_i + u_i$$

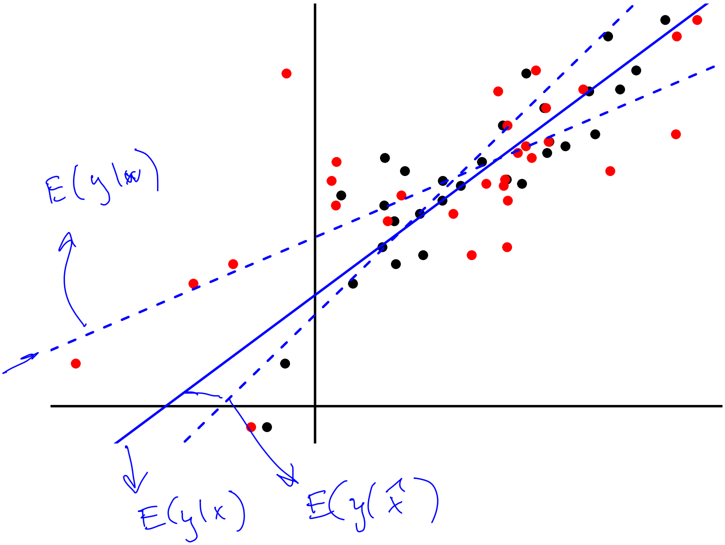
$$x_i = E(w_i)$$

- Find a model for $\boldsymbol{x} = \mathrm{E}(\boldsymbol{w}|\boldsymbol{z})$
- Replace unobserved $\boldsymbol{x}$ by $\hat{\mathrm{E}}(\boldsymbol{w}|\boldsymbol{z})$ in full model
- OLS variance estimates need correction (bootstrap or asymptotic)

Regression calibration

- Works a lot of the time
- Needs some model for the x
- Problematic if predictions aren't good, or if assumed model isn't good

Regression calibration



SIMEX

Simulation Extrapolation

$\hat{\beta}_1$

- Know that observed model estimate $\hat{\beta}_1^*$ is biased
- Try to get an idea of bias as ME cranks up
- Back track to estimate without bias

SIMEX

$w_i = x_i + u_i \qquad u_i \sim (0, \sigma_u^2)$

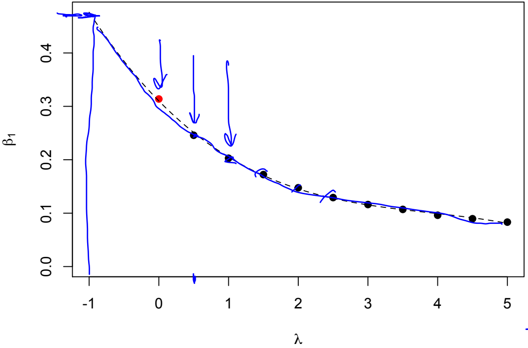
- Simulate new data $w_{b,i} = w_i + u_{b,i}$, where $u_{b,i} \sim \mathrm{N}\left[0, \lambda\sigma_u^2\right]$
- $\mathrm{Var}(w_{b,i}|x_i) = (1 + \lambda)\sigma_u^2$

  $1 - 1$
- Estimate $\hat{\beta}_1^*(\lambda)$
- Repeat many times, and for many $\lambda > 0$
- Extrapolate to $\lambda = -1$

SIMEX

- Computationally demanding
- Assumes you know (or have a good estimate of) σ_u^2
- Needs a good extrapolation model

SIMEX

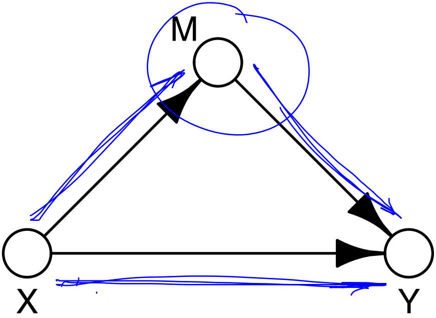


Mediation

- Mediation analyses attempt to understand mechanisms underlying data
- Specifically focused on assessing causation

What is mediation

- Predictor *x* influences outcome *y* through a *mediator m*



Assessing mediation

$\beta_{x,1} > \beta_{x,3}$

- Classical approach consists of three steps:
  - Regress y on x (total effect on outcome):

$$y_i = \beta_{0,1} + \boxed{\beta_{x,1}} x_i + \epsilon_i \qquad (1)$$

  - Regress m on x (direct effect on mediator)

$$m_i = \beta_{0,2} + \beta_{x,2} x_i + \epsilon_i$$

  - Regress y on x and m (direct and indirect effects on outcome)

$$y_i = \beta_{0,3} + \boxed{\beta_{x,3}} x_i + \beta_{m,3} m_i + \epsilon_i \qquad (3)$$

Assessing mediation

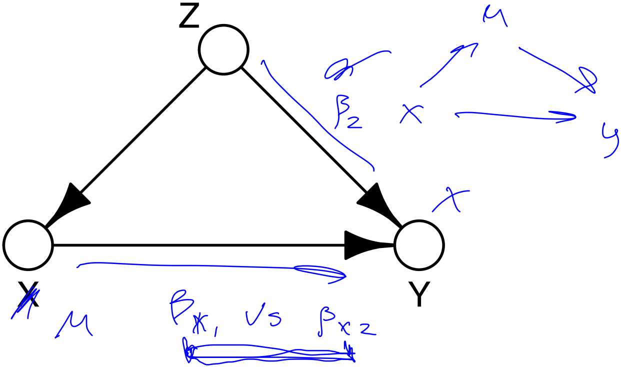
- To find a mediator
  - $\beta_{x,1}$ should be significant
  - $\beta_{x,2}$ should be significant
  - $\beta_{m,3}$ should be significant
- Typically $\beta_{x,3}$ is attenuated – closer to zero than $\beta_{x,1}$ – and is sometimes not significant
- $\beta_{x,2}\beta_{m,3}$ is often referred to as the indirect effect of x on y, and there are tests for this.

Declaring mediation?

- Mediation, conceptually, is about establishing causation
- Our data is often observational, and our regressions measure association
- Arguments for causation are often not statistical – biological plausibility, temporality, etc
- More recent work in causal inference is applicable but beyond this course

What is confounding

- Confounding occurs when the association between a predictor and outcome is distorted by a third variable
- Third variable z is associated with predictor x and outcome y; failing to adjust distorts association between x and y.



Assessing confounding

- To satisfy the conceptual definition:
  - Regress x on z (z is associated with x):
  - Regress y on z (z is associated with y)
  - Regress y on x (unadjusted association)

$$y_i = \beta_{0,1} + \beta_{x,1} x_i + \epsilon_i$$

  - Regress y on x and z (adjusted association)

$$\left[\; y_i = \beta_{0,2} + \beta_{\underline{x,2}} x_i + \beta_{z,2} z_i + \epsilon_i \;\right]$$

- (Remember that unadjusted associations are subject to omitted variable bias ...)

Assessing confounding

- There are similar rules for significance in confounding analysis
- Important confounders will be included regardless ✓
- This conceptual structure underlies much of what we've done in MLR

Mediation vs confounding

- Graphs are basically the same
- Analyses are basically the same
- Difference is conceptual, not statistical

Today's big ideas

- Measurement error, mediation, confounding
