

Precept 6: Regression

Soc 500: Applied Social Statistics

Ian Lundberg

Princeton University

October 20, 2016

Learning Objectives

- ① Review problem set 4
- ② Prepare for problem set 6
- ③ Other topics

Reviewing problem set 4

- ① Shay will explain problem 1.
- ② Aneesh will explain problem 2.
- ③ This one's good - good work all!
- ④ Hannah will explain problem 4.
- ⑤ Walk through estimators at the end of the answer key.

LaLonde training data

Data to evaluate the effects of a job training program

- re78 is earnings in 1978
- age is age
- educ is education, in years

Let's go through some examples in the R file.

Failing to include lower-order terms

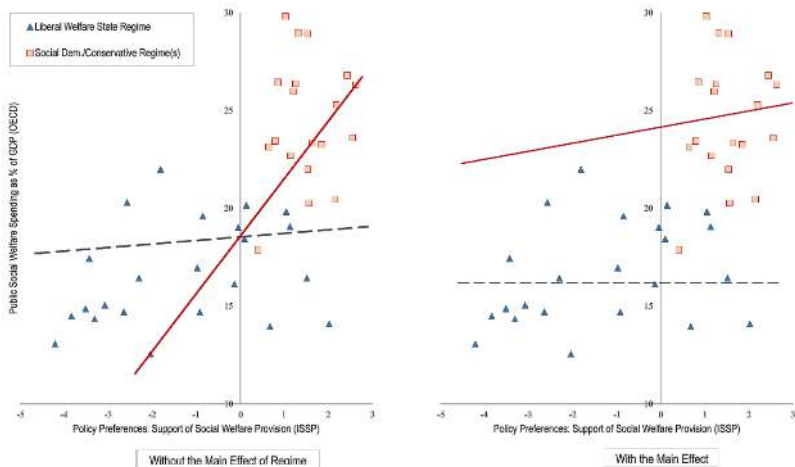


Figure 1: Predicted Regression Lines for the Effect of Policy Preferences on Social Welfare Spending, without and with the Main Effect of Regime

Attenuation bias: When X has random noise

What happens when X is measured with error?

$$\begin{aligned}\hat{\beta}_1 &= \frac{\text{Cov}(\tilde{X}, Y)}{\text{Var}(\tilde{X})} \\&= \frac{\text{Cov}(X + u, \beta X + \epsilon)}{\text{Var}(X + u)} \\&= \frac{\beta \text{Cov}(X, X) + \text{Cov}(X, \epsilon) + \text{Cov}(u, X) + \text{Cov}(u, \epsilon)}{\text{Var}(X) + \text{Var}(u) + 2\text{Cov}(X, u)} \\&= \beta \frac{\text{Var}(X) + 0 + 0 + 0}{\text{Var}(X) + \text{Var}(u) + 0} \\&= \beta \frac{\sigma_x^2}{\sigma_x^2 + \sigma_u^2} = \beta \frac{\sigma_x^2}{\sigma_{\tilde{x}}^2}\end{aligned}$$

$\hat{\beta}$ will thus be biased toward 0. We call this attenuation.

No bias when Y has random noise

What happens when Y is measured with error? No bias.

$$\begin{aligned}\hat{\beta}_1 &= \frac{\text{Cov}(X, \tilde{Y})}{\text{Var}(X)} \\&= \frac{\text{Cov}(X, \beta X + u + \epsilon)}{\text{Var}(X)} \\&= \frac{\beta \text{Cov}(X, X) + \text{Cov}(X, u) + \text{Cov}(X, \epsilon)}{\text{Var}(X)} \\&= \beta \frac{\text{Var}(X) + 0 + 0}{\text{Var}(X)} \\&= \beta\end{aligned}$$

Bigger standard error:

$$\widehat{SE}(\hat{\beta}_1) = \frac{\sigma^2}{\text{Var}(X)}$$

Weighting approach to regression

Brandon derived the OLS estimator for the slope as a weighted sum of the outcomes.

$$\hat{\beta}_1 = \sum_{i=1}^n W_i Y_i$$

Where here we have the weights, W_i as:

$$W_i = \frac{(X_i - \bar{X})}{\sum_{i=1}^n (X_i - \bar{X})^2}$$

This says that some observations have more *leverage* than others. I think we should talk through that intuition on the board.