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

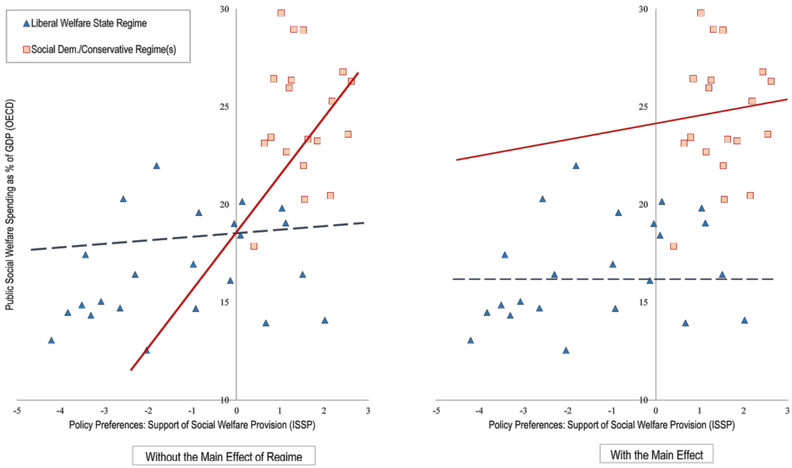

Figure 1: Predicted Regression Lines for the Effect of Policy Preferences on Social Welfare Spending, without and with the Main Effect of Regime

Attenuation bias: When X has random noise

What happens when X is measured with error?

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

$\hat{\beta}$ 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.

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

Bigger standard error:

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

Weighting approach to regression

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

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

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.