Precept 6: Regression

Soc 500: Applied Social Statistics

Alex Kindel

Princeton University

October 18, 2018

Today's plan[1]

1. Some homework items
   - PS4: Properties of variance
   - PS5: OLS assumptions
2. Dummy variables
3. Practice
   - Interpreting regression output
   - Partialing out
   - Interactions
   - Multicollinearity
4. If time...
   - Measurement error
   - `lm()` internals

[1] Thanks to Ian Lundberg and Brandon Stewart for some of today's examples and slides.

Properties of variance

Let $X_i \sim ?(\mu, \sigma^2)$.

$$\begin{aligned} Var[aX_j - bX_k] &= a^2 Var[X_j] + (-b)^2 Var[X_k] \\ &= a^2\sigma^2 + b^2\sigma^2 \end{aligned} \quad (1)$$

OLS assumptions

1. **Linearity in parameters**: We likely have omitted variable bias.

OLS assumptions

1. **Linearity in parameters**: We likely have omitted variable bias.
2. **Homoskedasticity**: The spread of the residuals increases in X.

OLS assumptions

1. **Linearity in parameters**: We likely have omitted variable bias.
2. **Homoskedasticity**: The spread of the residuals increases in X.
3. **Zero conditional mean**: Residuals are more negative for high values of X.

OLS assumptions

1. **Linearity in parameters**: We likely have omitted variable bias.
2. **Homoskedasticity**: The spread of the residuals increases in X.
3. **Zero conditional mean**: Residuals are more negative for high values of X.
4. **Random sampling**: FFCWS has a complex sample design; the unweighted observations are not quite representative (what's the population?).

Dummy variables

Today's example:

Today's example:

Today's example: **Pokèmon again!**

Just kidding.

Today's example: **LaLonde (1986)**

Famous[2] econometric analysis of 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.

[2] At this point in your education, you may have realized that academics should never be trusted when they refer to something as well-known.

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:

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