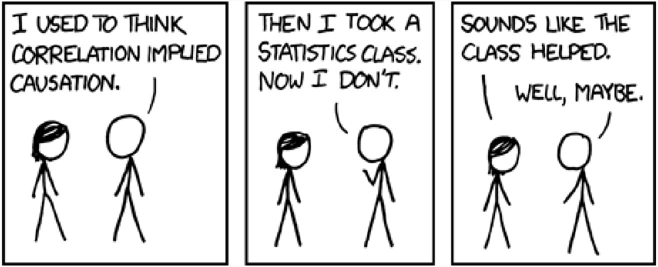
Impact Evaluation Methods

Topic 5: Regression

Alex Alekseev

University of Regensburg, Department of Economics


I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.

Previously on *Impact Evaluation Methods*...

- Stratification
- Matching
- Weighting

Regression and ATE

Simple Regression

- Suppose we are interested in the causal effect of a binary treatment variable D on the outcome variable Y
- We are considering a simple a regression model

$$Y = \alpha + \delta_R D + \epsilon.$$

- The regression coefficient in a simple binary regression is

$$\delta_R = \frac{Cov(Y, D)}{V(D)}.$$

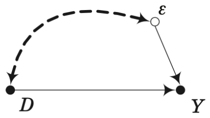
- It also equals the *NTE*:

$$\delta_R = \mathbb{E}[Y \mid D = 1] - \mathbb{E}[Y \mid D = 0].$$

Regression Adjustment

Confounders

- We might expect that D will be correlated with the error term in general



- In particular, there might be a set of variables X that confound the causal effect of D on Y

Model with Controls

- Suppose that X are observable and satisfy the conditional independence assumption
- We can include them in our regression model

$$Y = \alpha + \delta_R D + X'\beta + \epsilon^*$$

Omitted Variables Bias

- Let's pretend for a moment that we do not have X and run a naive regression of Y on D
- How bad would the bias be?
- The naive OLS coefficient from the regression of Y only on D is

$$\frac{Cov(Y,D)}{V(D)} = \delta_R + \beta' \rho_{XD}.$$

- This is the **omitted variables bias** formula
- Vector ρ_{XD} is the vector of regression coefficients of the components in X on D:

$$\rho_{XD} \equiv \left(\frac{Cov(X_1, D)}{V(D)}, ..., \frac{Cov(X_K, D)}{V(D)} \right)$$

Math time

- Consider the covariance between Y and D:

$$\begin{aligned} Cov(Y,D) &= Cov(\alpha + \delta_R D + X'\beta + \epsilon^*, D) \\ &= Cov(\alpha, D) + \delta_R Cov(D,D) + Cov(X'\beta, D) + Cov(\epsilon^*, D). \end{aligned}$$

- The first term is zero, since α is a constant
- The last terms are zero because we are assuming conditional independence
- The second term is simply $\delta_R V(D)$.
- Consider the third term

$$\begin{aligned} Cov(X'\beta, D) &= Cov(\beta_1 X_1, D) + ... + Cov(\beta_K X_K, D) \\ &= \beta_1 Cov(X_1, D) + ... + \beta_K Cov(X_K, D) \end{aligned}$$

- Then

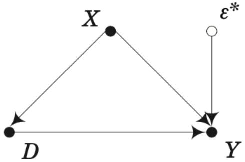
$$\frac{Cov(Y,D)}{V(D)} = \delta_R + \beta_1 \frac{Cov(X_1,D)}{V(D)} + ... + \beta_K \frac{Cov(X_K,D)}{V(D)} = \delta_R + \beta' \rho_{XD}.$$

OVB Explained

- The omitted variables bias formula says that our naive OLS coefficient would be unbiased if...
- ...the X variables are uncorrelated with D, $\rho_{XD} = 0$
- ...or if the X variables have no effect on Y or both
- However, if neither condition is true, X become **confounders**

Including Controls

- Suppose that we recognize the potential bias and include X in our model



- What would the regression coefficient δ_R represent?
- Is it the ATE?

Regression Anatomy

- Recall the **regression anatomy formula** (Frisch and Waugh, 1933)
- Suppose we have a model

$$Y = \beta_0 + \beta_1 X_1 + ... + \beta_K X_K + \epsilon.$$

- The kth regression coefficient in this formula is

$$\beta_k = \frac{Cov(Y, \tilde{X}_k)}{V(\tilde{X}_k)},$$

- where $\tilde{X}_k \equiv X_k - X'_{-k}\beta_{-k}$ is the residual from the regression of X_k on all other variables

Math time

- Consider the covariance between Y and $\tilde{X}_k$:

$$\begin{aligned} Cov(Y, \tilde{X}_k) = &Cov(\beta_0, \tilde{X}_k) + \beta_1 Cov(X_1, \tilde{X}_k) + ... \\ &+ \beta_k Cov(X_k, \tilde{X}_k) + ... + \beta_K Cov(X_K, \tilde{X}_k). \end{aligned}$$

- The term $Cov(\beta_0, \tilde{X}_k)$ is zero because β_0 is a constant
- The terms $\beta_j Cov(X_j, \tilde{X}_k), j \neq k$ are all zero because the residual $\tilde{X}_k$ will be uncorrelated with all other X.
- Consider the term $\beta_k Cov(X_k, \tilde{X}_k)$:

$$\beta_k Cov(X_k, \tilde{X}_k) = \beta_k Cov(\tilde{X}_k + X'_{-k}\beta_{-k}, \tilde{X}_k) = \beta_k V(\tilde{X}_k) + \beta_k Cov(X'_{-k}\beta_{-k}, \tilde{X}_k).$$

- The term $Cov(X'_{-k}\beta_{-k}, \tilde{X}_k) = 0$ since the residual $\tilde{X}_k$ is uncorrelated with all other X.
- Hence, we conclude that

$$\frac{Cov(Y, \tilde{X}_k)}{V(\tilde{X}_k)} = \beta_k.$$

Note

You can also write the regression anatomy formula as $\beta_k = \frac{Cov(\tilde{Y}, \tilde{X}_k)}{V(\tilde{X}_k)}$, where $\tilde{Y}$ is the residual from the regression of Y on all other variables except X_k. This is true since $Cov(\tilde{Y}, \tilde{X}_k) = Cov(Y - X'_{-k}\gamma, \tilde{X}_k) = Cov(Y, \tilde{X}_k) + Cov(X'_{-k}\gamma, \tilde{X}_k) = Cov(Y, \tilde{X}_k)$. The last equality follows from the fact that $\tilde{X}_k$ is uncorrelated with all other X.

Regression as Conditional-Variance-Weighted Matching

Fully Flexible Coding

- Recall our regression model in which we control for X

$$Y = \alpha + \delta_R D + X'\beta + \epsilon^*$$

- What kind of a treatment effect does δ_R represent?
- Suppose that we are adjusting for covariates using a **fully flexible coding**
- The fully flexible coding allows for a separate parameter for every value taken on by the control variables
- This model can be said to be **saturated-in-**X
- It is not fully saturated, however, because there are no interactions between D and X

Regression Coefficient as a Weighted Average

- The regression coefficient on D is a weighted average of conditional $ATE(X)$
- Weights are proportional to conditional variances of D (Angrist, 1998)

$$\delta_R = \mathbb{E}_X \left[ATE(X) \frac{V(D \mid X)}{\mathbb{E}_X V(D \mid X)} \right]$$

Math time

- From the regression anatomy formula, we get

$$\delta_R = \frac{Cov(Y, \tilde{D})}{V(\tilde{D})},$$

- where $\tilde{D}$ is the residual term in the regression of D on X
- Since the model is saturated-in-X, the conditional expectation function $\mathbb{E}[D \mid X]$ is linear in X and thus $\tilde{D} = D - X'\gamma = D - \mathbb{E}[D \mid X]$
- By the conditional expectation function (CEF) decomposition property, we also have that $\mathbb{E}[\tilde{D} \mid X] = 0$ and hence $\mathbb{E}\tilde{D} = \mathbb{E}_X \mathbb{E}[\tilde{D} \mid X] = 0$
- Therefore, $V(\tilde{D}) = \mathbb{E}\tilde{D}^2$ and $V(\tilde{D} \mid X) = \mathbb{E}[\tilde{D}^2 \mid X]$.

Math time

- Consider the variance of $\tilde{D}$

$$V(\tilde{D}) = \mathbb{E}\tilde{D}^2 = \mathbb{E}_X \mathbb{E}[\tilde{D}^2 \mid X] = \mathbb{E}_X V(\tilde{D} \mid X).$$

- On the other hand,

$$V(D \mid X) = V(\mathbb{E}[D \mid X] + \tilde{D} \mid X) = V(\tilde{D} \mid X),$$

- since $\mathbb{E}[D \mid X]$ is a constant after conditioning on X
- Hence,

$$V(\tilde{D}) = \mathbb{E}_X V(D \mid X).$$

Math time

- Now consider the covariance between Y and $\tilde{D}$:

$$Cov(Y, \tilde{D}) = \mathbb{E}[Y\tilde{D}] - \mathbb{E}Y\mathbb{E}\tilde{D} = \mathbb{E}[Y\tilde{D}],$$

$$\begin{aligned}
\mathbb{E}[Y\tilde{D}] &= \mathbb{E}[Y(D - \mathbb{E}[D \mid X])] \\
&= \mathbb{E}_X\mathbb{E}[Y(D - \mathbb{E}[D \mid X]) \mid X] \\
&= \mathbb{E}_X\left(\mathbb{E}[YD \mid X] - \mathbb{E}[Y\mathbb{E}[D \mid X] \mid X]\right) \\
&= \mathbb{E}_X\big(\mathbb{E}[Y^0D + (Y^1 - Y^0)D^2 \mid X] - \mathbb{E}[Y \mid X]\mathbb{E}[D \mid X]\big) \\
&= \mathbb{E}_X\big(\mathbb{E}[Y^0D \mid X] + \mathbb{E}[(Y^1 - Y^0)D^2 \mid X] - \mathbb{E}[Y \mid X]\mathbb{E}[D \mid X]\big) \\
&= \mathbb{E}_X\big(\mathbb{E}[Y^0 \mid X]\mathbb{E}[D \mid X] + \mathbb{E}[Y^1 - Y^0 \mid X]\mathbb{E}[D^2 \mid X] - \mathbb{E}[Y \mid X]\mathbb{E}[\\
&= \mathbb{E}_X\big(\mathbb{E}[D \mid X]\mathbb{E}[Y^0 - Y \mid X] + ATE(X)\mathbb{E}[D^2 \mid X]\big) \\
&= \mathbb{E}_X\big(\mathbb{E}[D \mid X](-\mathbb{E}[D(Y^1 - Y^0) \mid X]) + ATE(X)\mathbb{E}[D^2 \mid X]\big) \\
&= \mathbb{E}_X\big(ATE(X)\mathbb{E}[D^2 \mid X] - ATE(X)(\mathbb{E}[D \mid X])^2\big) \\
&= \mathbb{E}_X\big(ATE(X)(\mathbb{E}[D^2 \mid X] - (\mathbb{E}[D \mid X])^2)\big) \\
&= \mathbb{E}_X\left[ATE(X)V(D \mid X)\right].
\end{aligned}$$

Math time

- Therefore,

$$\delta_R = \frac{Cov(Y, \tilde{D})}{V(\tilde{D})} = \frac{\mathbb{E}_X\left[ATE(X)V(D \mid X)\right]}{\mathbb{E}_X V(D \mid X)} = \mathbb{E}_X\left[ATE(X)\frac{V(D \mid X)}{\mathbb{E}_X V(D \mid X)}\right].$$

- Hence, the regression coefficient on D is a weighted average of conditional $ATE(X)$ with weights being proportional to conditional variances of D

Expanding the Formula

- We can expand the formula for the regression coefficient

$$\delta_R = \sum_x ATE(x) \frac{V(D \mid X = x)\mathbb{P}(X = x)}{\sum_x V(D \mid X = x)\mathbb{P}(X = x)}$$

- Denote $p(X) = \mathbb{P}(D = 1 \mid X)$ to be the propensity score
- Recall that D is a Bernoulli random variable, then

$$\delta_R = \sum_x ATE(x) \frac{(1 - p(x))p(x)\mathbb{P}(X = x)}{\sum_x (1 - p(x))p(x)\mathbb{P}(X = x)}$$

Comparison to Treatment Effects

$$\delta_R = \sum_x ATE(x) \frac{(1 - p(x))p(x)\mathbb{P}(X = x)}{\sum_x (1 - p(x))p(x)\mathbb{P}(X = x)}$$

- This is different from our three treatment effects

$$ATE = \sum_x ATE(x)\mathbb{P}(X = x)$$

$$ATT = \sum_x ATE(x)\mathbb{P}(X = x \mid D = 1)$$

$$ATU = \sum_x ATE(x)\mathbb{P}(X = x \mid D = 0)$$

Comparison with ATT

- Make the following substitutions:

$$\begin{aligned}\mathbb{P}(X = x \mid D = 1) &= \frac{\mathbb{P}(D = 1 \mid X = x)\mathbb{P}(X = x)}{\mathbb{P}(D = 1)} \\ &= \frac{p(x)\mathbb{P}(X = x)}{\sum_x p(x)\mathbb{P}(X = x)}\end{aligned}$$

- Then ATT becomes

$$ATT = \sum_x ATE(x)\frac{p(x)\mathbb{P}(X = x)}{\sum_x p(x)\mathbb{P}(X = x)}$$

- Similarly, ATU becomes

$$ATU = \sum_x ATE(x)\frac{(1 - p(x))\mathbb{P}(X = x)}{\sum_x (1 - p(x))\mathbb{P}(X = x)}$$

Comparison with ATT and ATU

$$ATT = \sum_x ATE(x) \frac{p(x)\mathbb{P}(X = x)}{\sum_x p(x)\mathbb{P}(X = x)}$$

$$ATU = \sum_x ATE(x) \frac{(1 - p(x))\mathbb{P}(X = x)}{\sum_x (1 - p(x))\mathbb{P}(X = x)}$$

$$\delta_R = \sum_x ATE(x) \frac{(1 - p(x))p(x)\mathbb{P}(X = x)}{\sum_x (1 - p(x))p(x)\mathbb{P}(X = x)}$$

- ATT puts the most weight on covariate cells containing those who are most likely to be treated
- ATU puts the most weight on cells containing those who are most unlikely to be treated
- Regression puts the most weight on covariate cells where the conditional variance of treatment status is largest
- This variance is maximized when $p(X) = 0.5$

Why Does Regression Do That?

- Regression minimizes the **mean squared error**
- It gives more weight to stratum-specific effects with the **lowest expected variance**
- The expected variance of each stratum-specific effect is an **inverse** function of the stratum-specific variance of the treatment variable D

Comparison with ATE

$$ATE = \sum_x ATE(x)\mathbb{P}(X = x)$$

$$\delta_R = \sum_x ATE(x)\frac{(1 - p(x))p(x)\mathbb{P}(X = x)}{\sum_x (1 - p(x))p(x)\mathbb{P}(X = x)}$$

- Suppose that the propensity score is close to 0 or 1 for strata that have high total probability mass but close to .5 for strata with low probability mass
- Regression, under a fully flexible coding, can yield estimates that are far from the ATE even in an infinite sample

When Does Regression Identify ATE

$$ATE = \sum_x ATE(x)\mathbb{P}(X = x)$$

$$\delta_R = \sum_x ATE(x)\frac{(1 - p(x))p(x)\mathbb{P}(X = x)}{\sum_x (1 - p(x))p(x)\mathbb{P}(X = x)}$$

- Regressions would provide unbiased estimates of the *ATE* if either
  - the true propensity scores does not differ by strata or
  - the average stratum-specific causal effects do not vary by strata ($ATE(X)$ is constant)
- The first condition would imply that D is already independent of X

Linear Controls

- Under a **constrained** specification of X (e.g., in which some elements of X are constrained to have linear effects) the weighting scheme is more complex
- The weights remain a function of the marginal distribution of X and the stratum-specific conditional variance of D
- But the specific form of each of these components becomes conditional on the specification of the regression model (Angrist and Krueger, 1999)
- A linear constraint represents an implicit linearity assumption about true underlying propensity score that may not be linear in X

Fully Saturated Model

- Controlling for X as we did only helps to eliminate the **baseline bias** but not the **differential treatment effect bias**
- To eliminate the second type of bias, we would need to add all the **interactions** between X and D (**saturated** model)
- We would enact the same perfect stratification of the data as in matching
- None of the regression coefficients would immediately gives us the treatment effects we are looking for
- We would need to use the marginal distribution of X and the joint distribution of X given D to average the conditional treatment effects across the relevant distributions of X

Math time

- Recall our example

	$\mathbb{E}[Y^0 \mid D = 0, S]$	$\mathbb{E}[Y^1 \mid D = 1, S]$	$\mathbb{E}[\delta \mid S]$
$S = 1$	2	4	2
$S = 2$	6	8	2
$S = 3$	10	14	4

- Recall that $V(D \mid S) = p(S)(1 - p(S))$

Math time

$$\delta_R = \sum_x ATE(x) \frac{V(D \mid X = x)\mathbb{P}(X = x)}{\sum_x V(D \mid X = x)\mathbb{P}(X = x)}$$

	$S = 1$	$S = 2$	$S = 3$
$1 - p(S)$	$9/11$	$1/2$	$3/8$
$p(S)$	$2/11$	$1/2$	$5/8$
$V(D \mid S)$	$18/121$	$1/4$	$15/64$
$\mathbb{P}(S)$	0.44	0.24	0.32
$V(D \mid S)\mathbb{P}(S)$	0.065	0.06	0.075
weight	0.327	0.3	0.375

Math time

$$\delta_R = \sum_x ATE(x) \frac{V(D \mid X = x)\mathbb{P}(X = x)}{\sum_x V(D \mid X = x)\mathbb{P}(X = x)}$$

	$S = 1$	$S = 2$	$S = 3$
$1 - p(S)$	$9/11$	$1/2$	$3/8$
$p(S)$	$2/11$	$1/2$	$5/8$
$V(D \mid S)$	$18/121$	$1/4$	$15/64$
$\mathbb{P}(S)$	0.44	0.24	0.32
$V(D \mid S)\mathbb{P}(S)$	0.065	0.06	0.075
weight	0.327	0.3	0.375

Then the regression coefficient is

$$\delta = 2 \times 0.327 + 2 \times 0.3 + 4 \times 0.375 = 2.754.$$

Common Support

Common Support

- Neither regression nor matching give any weight to strata that do not contain both treated and control observations
- Consider a value of X say x^*, where either no one is treated or everyone is treated
- Then, $ATE(x^*)$ is undefined and the regression weights, $\mathbb{P}(D = 1 \mid X = x^*)(1 - \mathbb{P}(D = 1 \mid X = x^*))$, are zero
- Both regression and matching impose common support

Common Support in Regression

- Regression can make it easy to overlook these problems that are more explicit when doing matching
- Regression will implicitly drop strata for which the propensity score is either 0 or 1
- A researcher who interprets a regression result as a decent estimate of the ATT, but with supplemental conditional-variance weighting, may be entirely wrong
- No meaningful average causal effect may exist in the population

Example: Human Capital Revisited

- Consider the following the joint distribution $\mathbb{P}(D, S)$:

	$D = 0$	$D = 1$	$\mathbb{P}(S)$
$S = 1$	0.4	0	0.4
$S = 2$	0.1	0.13	0.23
$S = 3$	0.1	0.27	0.37
$\mathbb{P}(D)$	0.6	0.4	1

- The conditional distribution of S given D is

	$D = 0$	$D = 1$
$\mathbb{P}(S = 1 \mid D)$	2/3	0
$\mathbb{P}(S = 2 \mid D)$	1/6	0.325
$\mathbb{P}(S = 3 \mid D)$	1/6	0.675

Example: Human Capital Revisited

- Here are the corresponding potential outcomes (recall that we are assuming conditional independence)

	$\mathbb{E}[Y^0 \mid D=0, S]$	$\mathbb{E}[Y^1 \mid D=1, S]$	$\mathbb{E}[\delta \mid S]$
$S=1$	2	-	-
$S=2$	6	8	2
$S=3$	10	14	4

- The *ATT* can be estimated by considering only the values for those with S equal to 2 and 3:

$$ATT = 2 \times 0.325 + 4 \times 0.675 = 3.35$$

- There is no way to estimate the *ATU* and hence no way to estimate the *ATE*

Next Time on *Impact Evaluation Methods*...

Matching and regression practice