5 Regression

Regression and ATE

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

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

Recall that the regression coefficient in a simple binary regression is

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

It also equals the NTE:

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

Let's recall the conditions under which the regression coefficient δ_R identifies the ATE. Our starting point is the switching equation from the potential outcomes framework.

$$Y = Y^0 + (Y^1 - Y^0)D.$$

Denoting $\mu^0 \equiv \mathbb{E}Y^0$, $\mu^1 \equiv \mathbb{E}Y^1$, $v^0 \equiv Y^0 - \mathbb{E}Y^0$, and $v^1 \equiv Y^1 - \mathbb{E}Y^1$, and rearranging the terms yields

$$Y = \mu^0 + (\mu^1 - \mu^0)D + v^0 + (v^1 - v^0)D.$$

Notice that $\mu^1 - \mu^0$ is the ATE, and $v^0 + (v^1 - v^0)D$ can be thought of as the error term. It can also be rewritten as $Dv^1 + (1 - D)v^0$.

Now consider the NTE

$$\begin{aligned}
\mathbb{E}[Y \mid D = 1] - \mathbb{E}[Y \mid D = 0] &= \mu^1 - \mu^0 \\
&\quad + \mathbb{E}[v^0 + (v^1 - v^0)D \mid D = 1] - \mathbb{E}[v^0 + (v^1 - v^0)D \mid D = 0] \\
&= ATE \\
&\quad + \mathbb{E}[v^0 \mid D = 1] - \mathbb{E}[v^0 \mid D = 0] \\
&\quad + \mathbb{E}[(v^1 - v^0)D \mid D = 1] - \mathbb{E}[(v^1 - v^0)D \mid D = 0] \\
&= ATE + \mathbb{E}[v^0 \mid D = 1] - \mathbb{E}[v^0 \mid D = 0] + \mathbb{E}[(v^1 - v^0)D \mid D = 1].
\end{aligned}$$

Notice the similarity of this expression to the NTE decomposition we performed in 2 Potential Outcomes.

If the treatment effect is the same for everyone, $Y^1 - Y^0 = \delta = const$ (no differential treatment effect bias) and the expectation of v^0 does not depend on D, $\mathbb{E}[v^0 \mid D = 1] = \mathbb{E}[v^0 \mid D = 0]$ (no selection bias), then the regression coefficient δ_R is equal to the ATE. Notice that $Y^1 - Y^0 = \delta = const$ implies that $v^1 - v^0 = 0$.

These two conditions are equivalent to saying that the error term $v^0 + (v^1 - v^0)D$ is uncorrelated with D. On the flip side, if D is either correlated with v^0 (baseline bias) or

$(v^1 - v^0)D$ (differential treatment effect bias) or both, then D will be correlated with the error term and the regression coefficient δ_R will not identify the ATE.

Example

Suppose we have an equal number of people in the treatment and control groups, and within each group the outcomes are the same.

	Y^1	Y^0	D	Y	v^1	v^0	$v^0 + (v^1 - v^0)D$
Treatment group	20	10	1	20	5	5	5
Control group	10	0	0	0	-5	-5	-5

Since we have an equal number of people in each group, the expected values for each potential outcome are: $\mathbb{E}[Y^1] = 0.5 \times (20 + 10) = 15$ and $\mathbb{E}[Y^0] = 0.5 \times (10 + 0) = 5$. Using these values, we can compute v^1 and v^0, as well as the error term.

In this example, the treatment effect is the same for everyone (10), which means there is no differential treatment effect bias, hence $ATE = 10$. However, there is a baseline bias: v^0 is different across both groups. This leads to a correlation between D and the error term and ultimately leads to a bias in the NTE, which equals 20.

Consider a different scenario.

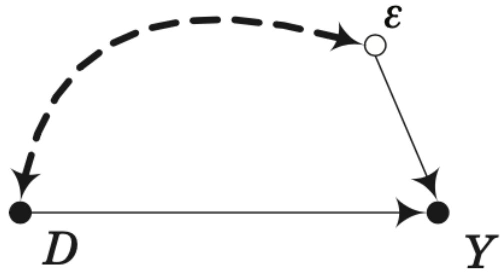
	Y^1	Y^0	D	Y	v^1	v^0	$v^0 + (v^1 - v^0)D$
Treatment group	20	10	1	20	2	0	2
Control group	16	10	0	10	-2	0	0

The expected values for each potential outcome are: $\mathbb{E}[Y^1] = 0.5 \times (20 + 16) = 18$ and $\mathbb{E}[Y^0] = 0.5 \times (10 + 10) = 10$.

There is no baseline bias: v^0 does not depend on D. However, there is a differential treatment effect bias: $ATT = 10$ while $ATU = 6$. The ATE is therefore 8. But the NTE is 10. It is biased because, as before, the error term is correlated with D. Although this time the reason for correlation is different.

Regression adjustment

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. Suppose that X are observable and satisfy the conditional independence assumption. We can thus include them in our regression model. Note that here X is a $K \times 1$ random vector, not a matrix.

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

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 OLS coefficient from a simple regression of Y 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 the regression coefficients of the components of X on D:

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

✓ **Math time: Proof of the omitted variables bias formula**

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: after controlling for X, D should not be correlated with the error term. The second term is simply $\delta_R V(D)$.

Consider the third term

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

Then

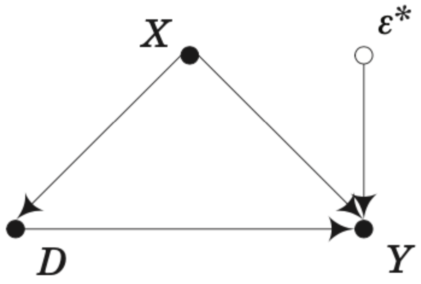
$$\frac{Cov(Y,D)}{V(D)} = \delta_R + \beta_1 \frac{Cov(X_1,D)}{V(D)} + \dots + \beta_K \frac{Cov(X_K,D)}{V(D)} = \delta_R + \beta' \rho_{XD}.$$

The omitted variables bias formula says that our naive OLS coefficient would be unbiased if

1. the X variables are uncorrelated with D, $\rho_{XD} = 0$, or
2. if the X variables have no effect on Y or both.

If neither condition is true, X become confounders.

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



What would the regression coefficient δ_R represent? What treatment effect would it correspond to? Before we answer this question, we need to recall the *regression anatomy formula* (Frisch and Waugh, 1933).

Suppose we have a model

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

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.

The regression anatomy formula says that the partial effect of variable X_k on Y is the slope coefficient in a regression of Y on the residualized X_k, i.e., X_k from which the effects of all other variables have been partialled out.

✓ **Math time: Proof of the regression anatomy formula**

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

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

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

Recall our regression model in which we control for X

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

Suppose that the specific way in which we are adjusting for covariates is by means of a *fully flexible coding*. In the fully flexible coding the variables are parameterized by an indicator variable for each possible combination of their values, save one for the reference category. The fully flexible coding allows for a separate parameter for every value combination taken on by the control variables. This model is said to be *saturated-in-*X. It is not fully saturated, however, because there are no interactions between D and X.

Math time: Conditional variance weighting

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

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,

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

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

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

since $\mathbb{E}\tilde{D} = 0$. Then we have

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

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 (Angrist, 1998).

We can expand the above expression

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

If we denote $p(X) = \mathbb{P}(D = 1 \mid X)$ to be the propensity score and recall that D is a Bernoulli random variable, then we have

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

This is different from our three treatment effects. Recall that

$$\begin{aligned}
ATE &= \sum_x ATE(x)\mathbb{P}(X = x) \\
ATT &= \sum_x ATE(x)\mathbb{P}(X = x \mid D = 1) \\
ATU &= \sum_x ATE(x)\mathbb{P}(X = x \mid D = 0).
\end{aligned}$$

The weights in these three formulas are given by either the marginal distribution of X or by the conditional distributions given D. In the regression case, the weights also include the conditional variance of D.

To make ATT and ATUm ore comparable with δ_R, 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, we can re-write ATU as

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

The ATT puts the most weight on covariate cells (strata) containing those who are most likely to be treated. The ATU puts the most weight on covariate cells (strata) containing those who are most unlikely to be treated. In contrast, regression puts the most weight on covariate cells where the conditional variance of treatment status is largest. As a rule, this variance is maximized when $p(X) = 0.5$, in other words, for cells where there are equal numbers of treated and control observations.

Why would regression implicitly invoke conditional-variance weighting as a supplement to weighting simply by, e.g., the marginal distribution of X? Regression minimizes the mean squared error. As a result, it gives more weight to stratum-specific effects with the lowest expected variance, and the expected variance of each stratum-specific effect is an inverse function of the stratum-specific variance of the treatment variable D. Thus, if the two pieces of the weighting scheme are not aligned (i.e., 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), then regression, even under a fully flexible coding of covariates, can yield estimates that are far from the true average treatment effect even in an infinite sample.

In general, regressions do not offer unbiased estimates of the average treatment effect when treatment effect heterogeneity is present, even under fully flexible coding. Regressions with fully flexible codings of the adjustment variables would provide unbiased estimates of the ATE if either (a) the true propensity scores does not differ by strata or (b) the average stratum-specific causal effects does not vary by strata ($ATE(X)$ is a constant). Notice that condition (a) would imply that D is already independent of X.

It is worth recalling that the weighting scheme for regression coefficient applies only under the fully flexible coding. 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 (see Angrist and Krueger 1999, Section 2.3.1). The basic intuition here is that a linear constraint represents an implicit linearity assumption about the true underlying propensity score that may not be linear in X.

Thus, controlling for X in the regression model, 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. Such a model would be called a *saturated* model. If a model is fully saturated (i.e., includes all the interactions between D and X in addition to main effects), then we would enact the same perfect stratification of the data as in matching. However, none of the regression coefficients would immediately give 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 in order to obtain the ATE, ATT, or ATU.

✓ Math time: Human capital revisited

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

We now complete the table to compute the regression weights for each conditional ATE. Recall that $V(D \mid S) = p(S)(1-p(S))$.

	$S=1$	$S=2$	$S=3$
$1-p(S)$	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

The weights obtain after we sum all the $V(D \mid S)\mathbb{P}(S)$ terms to get 0.2 and then divide each $V(D \mid S)\mathbb{P}(S)$ term by this sum.

Then the regression coefficient is

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

This number is different from all the treatment effects: $ATE = 2.64$, $ATT = 3$, $ATU = 2.4$. The regression coefficient is larger than the ATE, since regression upweights $ATE(S=3)$ and $ATE(S=2)$, for which the variance is relatively large, while downweighting $ATE(S=1)$, for which the variance is small, relative to the ATE that weights by $\mathbb{P}(S)$. On the other hand, the regression coefficient is lower than ATT that puts most weight on

$ATE(S=3)$ for which the propensity score is highest. Relative to ATT, the regression puts more weights on $ATE(S=2)$ for which the variance is highest.

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. That is, they are limited to strata where both treated and control observations are found.

However, 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. As a result, a researcher who interprets a regression result as a decent estimate of the average treatment effect, but with supplemental conditional-variance weighting, may be entirely wrong. No meaningful average causal effect may exist in the population.

Example: Common support in human capital

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

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

In this case, no individual for whom S is equal to 1 in the population is ever exposed to the treatment because $\mathbb{P}(S = 1, D = 1) = 0$. Because of this, the conditional expectation, $\mathbb{E}[Y \mid S = 1, D = 1]$ is undefined.

The ATT can be estimated as 3.35 by considering only the values for those with S equal to 2 and 3.

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

But there is no way to estimate the ATU (there is no counterfactual $\mathbb{E}[Y^0 \mid D = 1, S = 1]$ for $\mathbb{E}[Y^0 \mid D = 0, S = 1]$), and hence no way to estimate the ATE.

Homework

1. Derive the conditional distribution of S given D using the joint distribution.
2. Show that the NTE is equal to 8.05.