Outcome regression

1. What would the treated have experienced without treatment?

Suppose we observe whether each person received a treatment, their outcome, and some pre-treatment characteristics. Write these as D, Y, and X: $D = 1$ means treated, $D = 0$ means untreated, and X can contain several characteristics.

Each person has two potential outcomes: $Y(1)$ under treatment and $Y(0)$ without treatment. We observe only the one corresponding to their actual treatment:

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

The target in this note, as in the accompanying presentation, is the **average treatment effect on the treated (ATT)**:

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

Here $\mathbb{E}[\cdot \mid D = 1]$ means an average over the treated population. Split the difference into its two means:

$$\tau = \mathbb{E}[Y \mid D = 1] - \underbrace{\mathbb{E}[Y(0) \mid D = 1]}_{\theta}.$$

The first mean is observable: for treated people, $Y = Y(1)$. The second mean, which we call θ, is missing. **All the work is in finding θ.**

Why not use the ordinary control mean?

Controls reveal untreated outcomes. But their characteristics may differ from those of the treated. If higher-risk people are more likely to receive treatment, for example, the controls' average untreated outcome may be a poor guide to what the treated would have experienced.

Outcome regression keeps the observed control outcomes but reorganizes them before averaging. It first records, for each value of X, the mean outcome among controls with that value, and then averages those group means over the characteristics of the treated rather than those of the controls. Under an assumption explained below, that second average recovers θ.

There are two distinct questions to answer:

1. **Identification:** why does the reweighted average of control group means equal the missing mean?
2. **Consistency:** why does the estimate computed from data approach that population mean as the sample grows?

We first make the calculation concrete, then answer these questions in that order.

2. A small example: average the control means over the treated

Use the same example as the presentation. A characteristic X takes the values $0, 1, 2$, each in one third of the population. Treatment probabilities are $1/4, 1/2, 3/4$, respectively. We can express those population proportions as counts out of twelve. Suppose the mean untreated outcomes within the three groups are $0, 1, 4$, for both treated and control people, and that treatment adds $0, 2, 4$.

Characteristic x	Treated	Controls	Control mean outcome	Treated mean outcome
0	1	3	0	0
1	2	2	1	3
2	3	1	4	8
Total	6	6		

These are **population proportions expressed as counts**, not the promised outcome of a random sample of twelve people.

The ordinary control mean averages the control group means using control counts $(3, 2, 1)$:

$$\text{ordinary control mean} = \frac{3 \cdot 0 + 2 \cdot 1 + 1 \cdot 4}{6} = 1.$$

The missing untreated mean for the treated averages **the same three group means** using the treated counts $(1, 2, 3)$:

$$\theta = \frac{1 \cdot 0 + 2 \cdot 1 + 3 \cdot 4}{6} = \frac{7}{3}.$$

Only the distribution being averaged over has changed. The function being averaged, the control mean outcome at each x, is the same in both lines.

The observed treated mean is $(1 \cdot 0 + 2 \cdot 3 + 3 \cdot 8)/6 = 5$. Therefore,

$$\tau = 5 - \frac{7}{3} = \frac{8}{3} \approx 2.667.$$

The unadjusted difference would be $5 - 1 = 4$. Reweighting removes the part of that gap caused by the different mix of X.

The same number can be read one treated person at a time. Assign each treated person the control mean at their own value of x, and average the difference:

$$\frac{1 \cdot (0 - 0) + 2 \cdot (3 - 1) + 3 \cdot (8 - 4)}{6} = \frac{0 + 4 + 12}{6} = \frac{8}{3}.$$

This is the imputation reading: each treated person's counterfactual is the control mean in their covariate group, and the ATT is the average imputation error over the treated. Next we show why this works generally.

3. Why this particular function?

The **control regression function** is the mean observed outcome among controls at a given value of X:

$$\mu_0(x) = \mathbb{E}[Y \mid X = x, D = 0].$$

The subscript zero labels the true population function. Later, $\hat{\mu}(x)$ will denote its estimate. It is a feature of the observed data distribution: no counterfactual reasoning is needed to define it, and it can be computed from the control sample alone.

When can this recover an untreated outcome for the treated?

Averaging over the treated composition is only part of the argument. We also need controls to provide the right untreated mean *within* each covariate group.

Comparable untreated outcomes given X. Assume $Y(0) \perp\!\!\!\perp D \mid X$: once X is held fixed, treatment status conveys no further information about the untreated potential outcome. This is conditional unconfoundedness. In particular, for groups represented among the treated,

$$\mathbb{E}[Y(0) \mid X = x, D = 1] = \mathbb{E}[Y(0) \mid X = x, D = 0] = \mathbb{E}[Y \mid X = x, D = 0] = \mu_0(x).$$

The first equality uses the causal assumption; the second uses $Y = Y(0)$ for controls. Equality of these conditional means is all the proof needs. Fitting a regression cannot establish that assumption from the observed data.

Available controls. Write $\pi = \mathbb{P}(D = 1)$ and $p_0(x) = \mathbb{P}(D = 1 \mid X = x)$, and assume $\pi > 0$ and $p_0(X) \leq 1 - \epsilon$ for some $\epsilon > 0$. This overlap condition ensures that $\mu_0(x)$ is defined wherever there are treated people. Values with $p_0(x) = 0$ are allowed: they contain no treated people, and μ_0 is never evaluated there. Note the asymmetry of the construction: the function is fitted where the controls are and evaluated where the treated are.

For the expectations and convergence arguments below, assume independent, identically distributed observations and finite absolute means of both potential outcomes. These are sampling and moment conditions, separate from the causal assumption.

4. The population proof: the control regression recovers the missing mean

The treatment indicator D will do the work of restricting attention to the treated. It is one for treated people and zero for controls, so multiplying by it discards the controls and leaves the treated counted at their population share.

First find the total weight

Given X, the mean of D is the treatment probability, $\mathbb{E}[D \mid X] = p_0(X)$. The law of iterated expectations says that averaging conditional means gives the overall mean, $\mathbb{E}[\mathbb{E}[A \mid X]] = \mathbb{E}[A]$. Consequently,

$$\mathbb{E}[D] = \mathbb{E}[\mathbb{E}[D \mid X]] = \mathbb{E}[p_0(X)] = \pi.$$

The average treatment probability is the treated share.

Then find the imputed outcome contribution

Once X is fixed, $\mu_0(X)$ is a fixed number, so it factors out of the conditional expectation:

$$\mathbb{E}[D\,\mu_0(X) \mid X] = \mu_0(X)\mathbb{E}[D \mid X] = p_0(X)\mu_0(X).$$

Averaging over X gives

$$\mathbb{E}[D\,\mu_0(X)] = \mathbb{E}[p_0(X)\mu_0(X)].$$

Why does this concern the treated? For covariate groups with $p_0(X) > 0$,

$$\begin{aligned}\mathbb{E}[Y(0) \mid X, D = 1] &= \mathbb{E}[Y(0) \mid X, D = 0] \\ &= \mathbb{E}[Y \mid X, D = 0] = \mu_0(X).\end{aligned}$$

The first equality uses unconfoundedness; the second uses $Y = Y(0)$ for controls. Thus the treated fraction times its mean untreated outcome is

$$\begin{aligned}\mathbb{E}[DY(0) \mid X] &= p_0(X)\mathbb{E}[Y(0) \mid X, D = 1] \\ &= p_0(X)\mu_0(X).\end{aligned}$$

Where $p_0(X) = 0$, both contributions are zero and no treated mean is needed. Averaging over X now gives

$$\begin{aligned}\mathbb{E}[p_0(X)\mu_0(X)] &= \mathbb{E}[\mathbb{E}[DY(0) \mid X]] \\ &= \mathbb{E}[DY(0)] \\ &= \pi\mathbb{E}[Y(0) \mid D = 1] = \pi\theta.\end{aligned}$$

The indicator D retains the treated people's untreated potential outcomes and assigns zero to everyone else; their overall contribution is their population share times their group mean.

Divide to obtain a mean

We have shown that the imputed outcome contribution is $\pi\theta$ and the total weight is π. Since $\pi > 0$,

$$\boxed{\frac{\mathbb{E}[D\,\mu_0(X)]}{\mathbb{E}[D]} = \frac{\pi\theta}{\pi} = \theta, \qquad \text{equivalently} \qquad \theta = \mathbb{E}[\mu_0(X) \mid D = 1].}$$

This is identification: a function of observed data equals the missing mean.

The middle quantity $\mathbb{E}[p_0(X)\mu_0(X)]$ is worth noticing. The companion note on inverse probability weighting arrives at the same expression from the other side, by showing that odds-weighted control outcomes satisfy $\mathbb{E}[(1-D)o_0(X)Y] = \mathbb{E}[p_0(X)\mu_0(X)]$. The two representations of θ agree at the truth because both equal this one population expectation.

5. From a population identity to something we can calculate

In data, replace the population means by sample averages and the unknown control means by estimates. For observations $i = 1, \dots, n$:

1. Estimate $\hat{\mu}(x)$, the mean outcome among controls with characteristics x, using the control observations.
2. Predict $\hat{\mu}(X_i)$ for each treated observation.
3. Average the differences $Y_i - \hat{\mu}(X_i)$ over the treated.

Writing $n_1 = \sum_{i=1}^{n} D_i$ for the number treated, the estimator is

$$\boxed{\hat{\tau} = \underbrace{\frac{\sum_{i=1}^{n} D_i Y_i}{n_1}}_{\text{treated mean}} - \underbrace{\frac{\sum_{i=1}^{n} D_i \hat{\mu}(X_i)}{n_1}}_{\text{imputed control mean } \hat{\theta}} = \frac{1}{n_1} \sum_{i=1}^{n} D_i (Y_i - \hat{\mu}(X_i)).}$$

This requires at least one treated observation. Both terms have the same denominator, the treated count, so the estimator is an ordinary average over the treated of an observed number and an imputed one. Nothing is divided by an estimated probability, so no weight can explode; what replaces that risk is the risk of extrapolation, discussed in Section 8.

For example, with treated outcomes $0, 3, 3, 8, 8, 8$ and imputations $0, 1, 1, 4, 4, 4$, the estimate is $(0 + 2 + 2 + 4 + 4 + 4)/6 = 8/3$, which is the per-person calculation of Section 2.

Why fit the regression on the controls only?

A pooled regression of Y on a constant, D, and a full set of indicators for X would estimate a single coefficient on D, imposing one effect for all covariate groups. When the effect varies with x, that coefficient is a weighted average of the group-specific effects $\delta(x) = \mathbb{E}[Y(1) - Y(0) \mid X = x]$, with weights proportional to $p_0(x)(1 - p_0(x))\mathbb{P}(X = x)$, the conditional variance of treatment, rather than weights proportional to the treated share $p_0(x)\mathbb{P}(X = x)$. In the example the three groups are equally common and the effects are $\delta = (0, 2, 4)$, so the variance weights are proportional to $(3/16, 4/16, 3/16)$ and the pooled coefficient is

$$\frac{3 \cdot 0 + 4 \cdot 2 + 3 \cdot 4}{10} = 2 \neq \frac{8}{3} = \tau.$$

The two differ because the pooled regression weights groups by how evenly treatment is split within them, whereas the ATT weights them by how many treated people they contain. Fitting μ_0 on the controls and averaging over the treated avoids imposing a common effect.

Why not also model the treated outcomes?

For the ATT the treated mean is observed, so only μ_0 needs to be modelled. The average effect in the *whole* population (ATE) requires both $\mathbb{E}[Y \mid X, D = 1]$ and $\mathbb{E}[Y \mid X, D = 0]$, and therefore offers a second opportunity for misspecification. That is a different estimand, not a refinement of this one.

6. Consistency when the true regression function is known

Consistency means $\hat{\tau} \xrightarrow{p} \tau$: for any fixed tolerance, the probability that the estimation error exceeds that tolerance tends to zero as n grows. It does not mean the estimate equals τ in a particular sample.

First imagine that we know μ_0 and impute with it directly. This isolates ordinary sampling error from the error in estimating the regression function.

Write the estimator as a ratio of sample averages

Divide its numerator and denominator by n:

$$\hat{\tau}_{\text{true } \mu} = \frac{n^{-1} \sum_i D_i(Y_i - \mu_0(X_i))}{n^{-1} \sum_i D_i}.$$

Apply the law of large numbers to each

The **law of large numbers** says a sample average of i.i.d. observations with a finite absolute mean converges in probability to its expectation. Apply it twice:

$$\frac{1}{n} \sum_i D_i(Y_i - \mu_0(X_i)) \xrightarrow{p} \mathbb{E}[D(Y - \mu_0(X))], \qquad \frac{1}{n} \sum_i D_i \xrightarrow{p} \pi.$$

The first expectation is exactly what Section 4 evaluated. Splitting it,

$$\mathbb{E}[DY] - \mathbb{E}[D\mu_0(X)] = \pi \mathbb{E}[Y \mid D = 1] - \pi\theta = \pi\tau,$$

where the first term uses the same indicator argument as before, now applied to the observed outcome.

Divide the limits

Because the denominator approaches the positive number π, the ratio approaches the ratio of the limits:

$$\hat{\tau}_{\text{true } \mu} \xrightarrow{p} \frac{\pi\tau}{\pi} = \tau.$$

This is the bridge: **identification gives the correct population limit; the law of large numbers takes the sample estimator to that limit.** We must still show that estimating the regression function does not change it.

7. Why estimating the regression function need not change the limit

Fitted values use the data, so we cannot simply treat the imputations as i.i.d. observations. Instead, compare the fitted-function sums with the true-function sums we have already proved converge.

Here is a transparent **sufficient condition**. Suppose the largest imputation error tends to zero:

$$r_n := \sup_x |\hat{\mu}(x) - \mu_0(x)| \xrightarrow{p} 0,$$

where the supremum runs over the covariate support. This is uniform consistency: eventually, even the largest error is small with high probability. It is stronger than necessary, but lets us give a short, complete proof, including when the regression is fitted on the same sample.

Small function errors imply small errors in the sample averages

Only the imputation term involves $\hat{\mu}$, and the treatment indicator is bounded by one. Therefore,

$$\left|\frac{1}{n}\sum_i D_i\hat{\mu}(X_i) - \frac{1}{n}\sum_i D_i\mu_0(X_i)\right| \leq r_n \frac{1}{n}\sum_i D_i \leq r_n \xrightarrow{p} 0.$$

The bound holds for each realized dataset, without requiring independence between $\hat{\mu}$ and the observations it is evaluated at. The denominator $n^{-1}\sum_i D_i$ does not involve $\hat{\mu}$ at all, so it is unaffected.

Estimated-function averages therefore have the same limits as true-function averages: the numerator approaches $\pi\tau$ and the denominator approaches $\pi > 0$. Their ratio approaches τ, so $\hat{\tau} \xrightarrow{p} \tau$.

Where the overlap condition went

Compare this with the corresponding step for inverse probability weighting, where the weights must be kept away from the region in which p_0 approaches one before the analogous bound can be written. No such restriction appears here, because the multiplier D_i is at most one. Overlap has not become unnecessary; it entered earlier, in Section 3, as the condition under which $\mu_0(x)$ is defined and estimable at the covariate values the treated actually have. Weighting makes a violation visible in the arithmetic, as an exploding weight. Outcome regression absorbs it silently into an extrapolated prediction.

The uniform condition is also stronger than necessary. The bound above only needs the average absolute error over the treated covariate distribution to vanish, which is implied by $\|\hat{\mu}-\mu_0\|_2 \xrightarrow{p} 0$ together with a sample-splitting or complexity restriction on the fitted function. The supremum is used here because it requires neither.

8. Making the conditions concrete

In the example, estimate the control means by group averages

With $X \in \{0, 1, 2\}$, let n_{0x} be the number of controls in group x and $\bar{Y}_{0x}$ their mean outcome. Estimate

$$\hat{\mu}(x) = \bar{Y}_{0x} = \frac{\sum_i (1 - D_i)\, \mathbf{1}\{X_i = x\}\, Y_i}{\sum_i (1 - D_i)\, \mathbf{1}\{X_i = x\}},$$

where $\mathbf{1}\{\cdot\}$ is one when its condition holds and zero otherwise, so that both sums run over the controls in group x alone. Divided by n, the two sums are sample averages. They converge to $\mathbb{P}(X = x, D = 0)\mu_0(x)$ and to $\mathbb{P}(X = x, D = 0) = \mathbb{P}(X = x)(1 - p_0(x))$, which overlap and $\mathbb{P}(X = x) > 0$ make positive. Their ratio therefore converges to $\mu_0(x)$. There are only three groups, so the maximum of their three absolute estimation errors also approaches zero: this gives uniform consistency. Groups with treated people but no controls prevent the calculation, and with finitely many positive-probability groups that problem disappears with probability tending to one.

With this saturated estimate, the imputed mean has an exact form. Writing n_{1x} for the treated count in group x,

$$\hat{\theta} = \frac{\sum_x n_{1x} \bar{Y}_{0x}}{\sum_x n_{1x}}.$$

This is the same expression that the companion note obtains for the frequency-based inverse probability weighting estimator. When the covariate is discrete and both nuisance functions are estimated without restriction, the two estimators are numerically identical in every sample. They part company only once a model restricts one of them, which is the situation the presentation on doubly robust estimation takes up.

What the result does and does not promise

The regression estimates must be adequate. A fitted model does not automatically converge to μ_0. If it converges to a wrong function μ^*, the imputed mean, under analogous convergence conditions, approaches $\mathbb{E}[\mu^*(X) \mid D = 1]$, and the resulting bias is

$$\mathbb{E}[\mu_0(X) - \mu^*(X) \,\big|\, D = 1] = \frac{1}{\pi}\mathbb{E}[p_0(X)(\mu_0(X) - \mu^*(X))].$$

The bias is linear in the function error, undampened, and averaged against the *treated* covariate distribution: an error in the region where the treated are concentrated passes straight through.

The example makes its size concrete. Write $q = (1/2, 1/3, 1/6)$ for the control covariate distribution and fit a line by least squares under it:

$$b = \frac{\mathrm{Cov}_q(X, \mu_0(X))}{\mathrm{Var}_q(X)} = \frac{1}{5/9} = \frac{9}{5}, \qquad a = 1 - \frac{9}{5} \cdot \frac{2}{3} = -\frac{1}{5},$$

so $\mu^\dagger = (-0.2, 1.6, 3.4)$ against a truth of $(0, 1, 4)$, with errors $(1/5, -3/5, 3/5)$. Averaged over the treated shares $(1/6, 2/6, 3/6)$ the bias is $2/15 \approx 0.133$, and the estimand becomes $8/3 + 2/15 = 2.8$. The errors fail to cancel because half the treated mass sits at $x = 2$, where the line understates μ_0 by 0.6, while the offsetting error at $x = 1$ carries a third of the mass. A fit that looks acceptable on the control sample, whose mass sits at $x = 0$, is then evaluated where the control sample is thin.

Overlap and the causal assumption still matter. Where $p_0(x)$ is close to one, few controls are available at x and the prediction there rests on the functional form rather than on data. The estimator still returns a finite, smooth number in that case, which is a reason for caution rather than reassurance. Unmeasured differences in untreated outcomes within X groups cannot be repaired by any choice of regression function.

The argument requires both links: the assumptions make control group means the treated counterfactual, and accurate estimation of those means plus averaging over the treated makes the sample estimate approach it. No finite-sample unbiasedness is claimed.