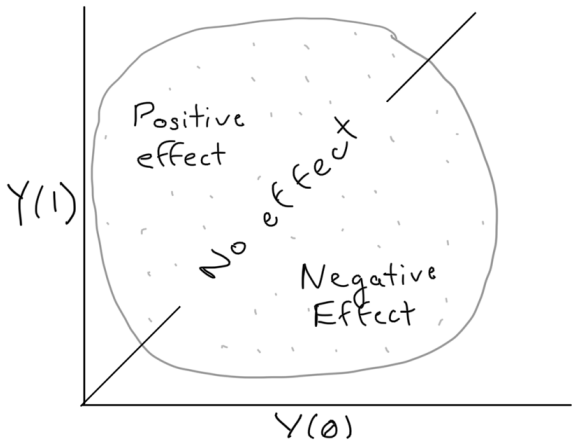
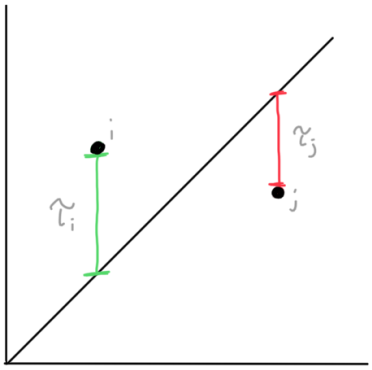
Visualizing the Potential Outcomes, Y(0) and Y(1)

One place we might begin is by supposing that we knew the full science table and could visualize it graphically. In reality, the closest we may ever get to this in an applied space is by fitting models for the expected outcome under the treatment, and control assignments. However, in that scenario, the model for control outcomes (prognostic score) will always be extrapolating to the treatment group, and vice-versa for the control group.

However, supposing we had the complete science table at our disposal, we could visualize individual treatment effects spatially:



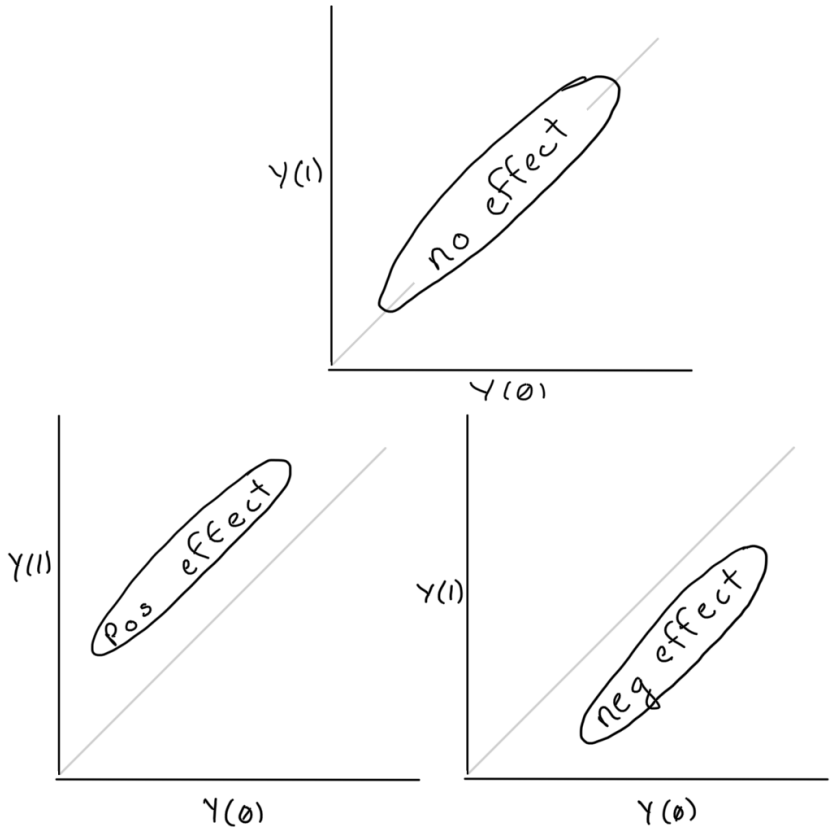
An individual's treatment effect is visualized by the difference between their horizontal and vertical coordinates. Individuals on the diagonal have no treatment effect. Individuals above and below the diagonal have positive and negative treatment effects, respectively, and a greater distance from the diagonal corresponds to a greater response to treatment. An interesting observation is the individual treatment effect for an observation is the vertical (or, equivalently, the horizontal) distance to the diagonal. These look a little like residuals:



Some things to note:

- I'm showing just the first quadrant above, which is often - though not necessarily - the quadrant in which the potential out comes will lie.
- Implicitly, Y is a continuous outcome, above. Visualizing a binary outcome can be less interesting, although one might consider visualizing the probabilities of the positive outcome under the treatment or control assignment instead. This may or may not make sense depending on what you consider to be fixed or random in your probabilistic framework.

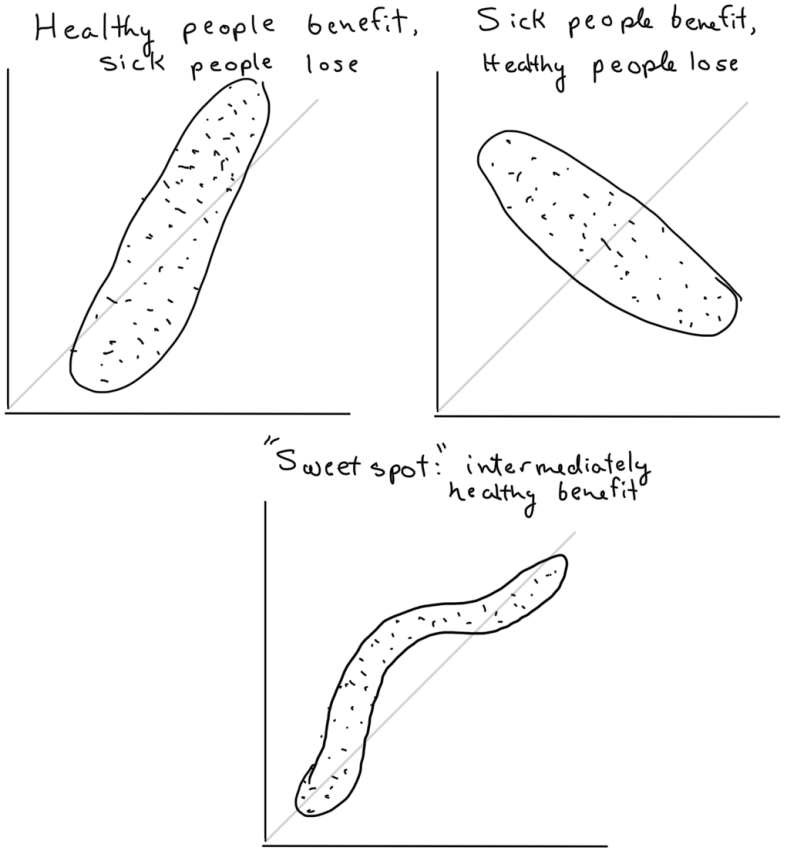
The shape of the data cloud in this space visually indicates the treatment effect. For example, a constant additive effect might be visualized as:



If the treatment effect is additive, the data cloud must be parallel to the diagonal. Any deviation from the diagonal indicates treatment effect heterogeneity. A multiplicative treatment effect, for example, might appear as a data cloud which takes a quadratic shape.

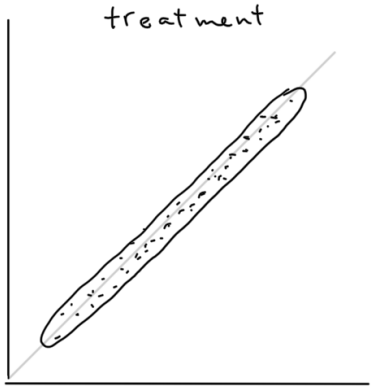
Some Visualizations of Different Treatment Effect Possibilities

The sketches below visualize some other possible treatment effects. Suppose for consistency that a more positive outcome is more desirable, so people with a low Y might be thought of as "sicker" in medical applications, for example, and those with a high Y might be thought of as "healthier."

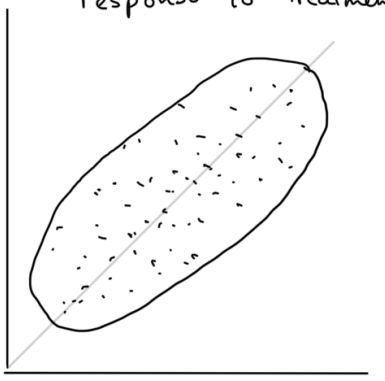


Another consideration is the dispersion of the point cloud. A more dispersed point cloud indicates greater variation in treatment effect. The dispersion may also be correlated with the potential outcomes, as shown below:

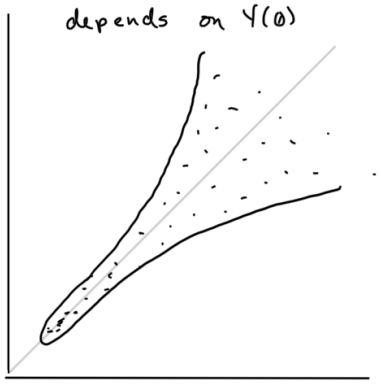
Uniform response to treatment



Heterogeneous response to treatment.



Variation in Treatment Respose depends on $Y(0)$



A Note on Applications

As always, the fundamental problem of causal inference is that the science table is not known. Some causal inference methods rely on very sophisticated modeling of the potential outcomes in order to estimate the unobserved potential outcomes. However, it is necessarily true that a model for the control outcome must extrapolate to the treatment group, and the same is true for a model of the treated outcome in the treatment group. Fundamental to matching approaches is an underlying skepticism of the models we fit - for outcome or treatment - because this is what justifies matching in the first place than adjusting for confounding directly in the analysis phase.

In practice, one could fit models for the potential outcomes and visualize the resulting estimated outcomes on the remainder of the sample. This could be used to suggest a potential model for treatment effect heterogeneity. However, any examination of these plots should be treated as data-exploration, not confirmatory research.

Matching

Should you match on predicted treatment outcomes? Interestingly, maybe not. If the estimand is the sample average treatment effect among the treated, then the control potential outcomes under the treatment may not matter. As an illustration, let $\{Y_i(0), Y_i(1)\}_{i=1}^{n_1}$ represent the potential outcomes of each of the treated individuals in a sample and $\{Y_{i'}(0), Y_{i'}(1)\}_{i=1}^{n_1}$ represent the potential outcomes of each of n_1 matched control individuals. The sample average treatment effect among the treated is:

$$\tau^{SATT} = \sum_{i=1}^{n_t} Y_i(1) - Y_i(0)$$

If we estimate this by

$$\hat{\tau}^{SATT} = \sum_{i=1}^{n_t} Y_i(1) - Y_i'(0)$$

Then our error is:

$$\tau^{SATT} - \hat{\tau}^{SATT} = \sum_{i=1}^{n_t} Y_i(0) - Y_i'(0)$$

This doesn't actually depend on how close $Y_{i'}(1)$ is to $Y_i(1)$. Inessence, since we only care about the control individuals as proxies for understanding the control potential outcomes of the treated individuals, the unobserved control potential outcomes, $\{Y_i'(1)\}_{i=1}^{n_t}$ were inconsequential. Of course, this might change if you were more interested in a more symmetric estimand, such as the ATE.