MS&E 228: Class Polls Winter 2025

Vasilis Syrgkanis

Asst. Professor, MS&E

When we run a linear regression y = beta'X + epsilon. Do we expect the residual of the prediction epsilon to satisfy: E[epsilon | X] = 0?

69

Answer

- We can always expect it to be uncorrelated

$$E[\epsilon \cdot X] = 0$$

- However, for the conditional expectation to hold, we need the linear model to be well-specified, i.e.

$$E[\, Y \mid X \,] = \beta' X$$

I have data Y, D, W from an experiment, with Y outcome, D treatment and W covariates. If I run OLS of Y on 1, D, W, then the coefficient associated with D is not the average treatment effect if the relationship of Y with D, W is not linear.

Answer

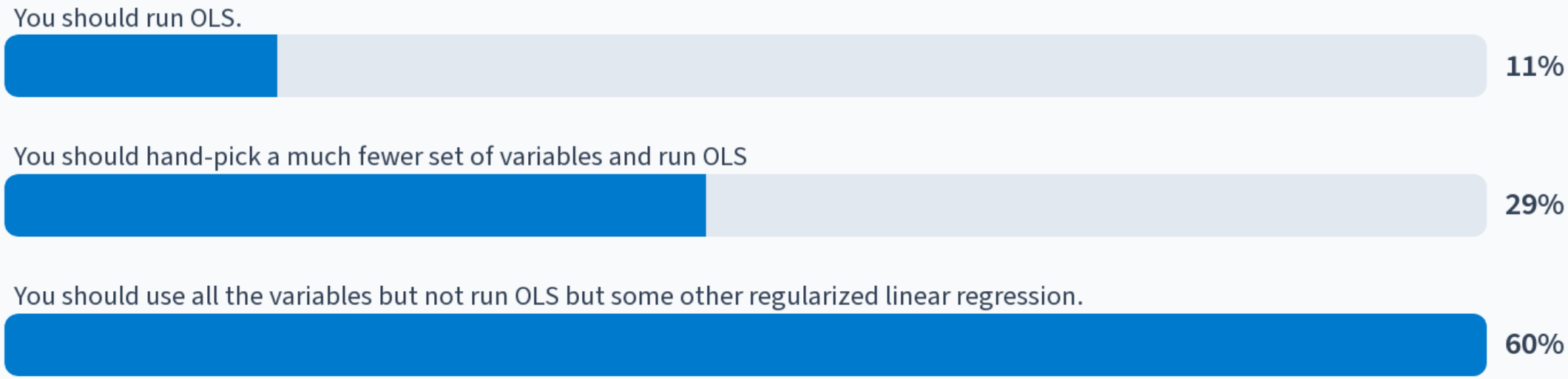
- For randomized control trial data, the coefficient associated with D in such a regression will converge to the ATE, even if the conditional expectation is not linear and the model is not well-specified.

If I have data from a randomized experiment and I run OLS Y on D, 1, W, then the precision of estimate the ATE gets better the more variables W I add (assuming number of variables is much smaller than number of samples).

Answer

- Not necessarily
- It does so if the model is well-specified
- It also does so if I do an interactive regression were I also control for the interactions with de-meaned covariates W
- But just adding controls, there is a slight risk that variance of the ATE can increase
- Typically it does not

If you have data from an RCT with 400 samples and 10 covariates. One covariate is categorical taking 20 values. You want to control for all pairwise interactions of the other 9 variables with the one hot encoding of the categorical.



Answer

- You can use double lasso techniques to control for high-dimensional sets of confounders and don’t need to hand-pick them.

If I run Lasso to calculate my best linear predictor, I can construct correct confidence intervals by resampling the data with replacement (bootstrap), refitting the lasso and calculating the standard deviation of the refitted estimates.

Answer

- You should not bootstrap the plain lasso
- Lasso is biased and bootstrapping will not capture the bias
- You can bootstrap the double lasso procedure if you want

If I run post lasso OLS, then I can look at the confidence intervals that are calculated in the OLS phase and use them for statistical significance.

Answer

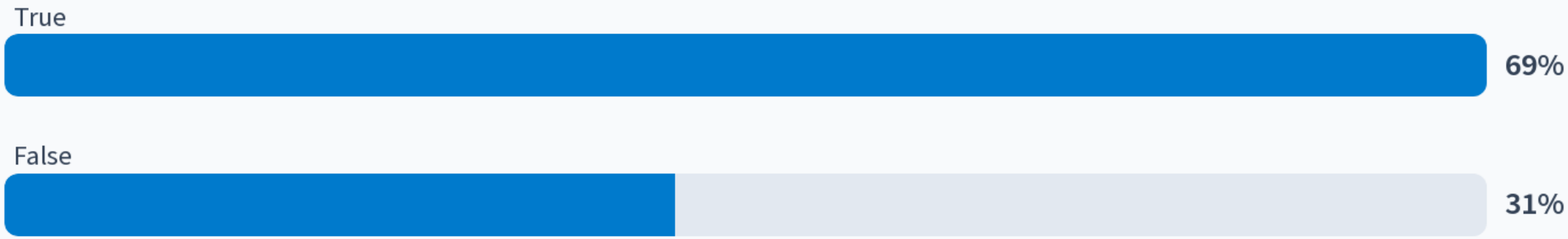
- No. These confidence intervals do not account for the selection step
- You can do the “double selection procedure” instead if you want
- But you need to include the controls chosen both from the lasso of Y on controls and of D on controls.
- But not just the Y on controls features.

I have data that do not stem from a randomized trial. They contain an outcome of interest Y, a treatment D and high-dimensional covariates X. The coefficient recovered by the double lasso recovers average causal effect of D.

Answer

- It is not clear that there is no omitted confounder

A domain expert asserted to me that the decision to administer the treatment was only decided based on X (and maybe other random factors irrelevant to the outcome). The coefficient returned by the double lasso recovers average causal effect of D.



Answer

- It is now clear that there is no omitted confounder
- However, the double lasso makes linearity assumptions
- The coefficient associated with D, will converge to the ATE only if the true CEF is well-specified, i.e. conditional expectation of Y on X is linear
- Unlike the RCT case (were we did not need well-specification), there is no guarantee otherwise that the coefficient associated with D will converge to the ATE

You have access to observational data containing Y, D, and X. How would you judge if conditional exogeneity holds, conditional on some subset of the variables X?

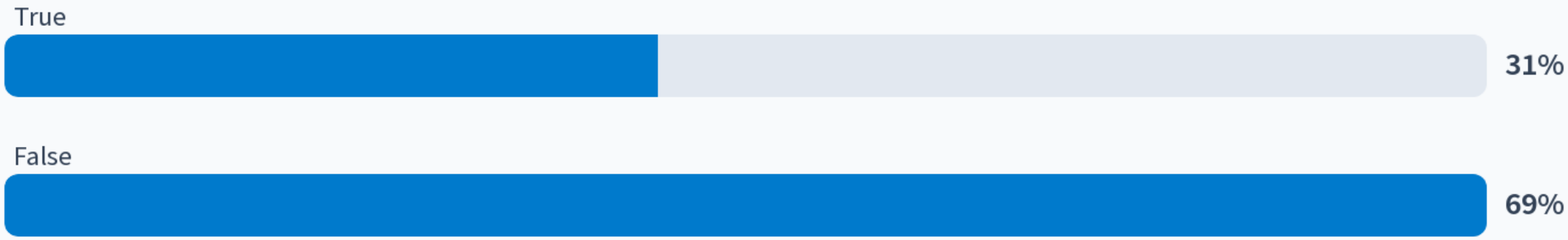
I would check if (Y(1),Y(0)) is independent of D|X

SEE MORE

Answer

- There is no fully data driven approach to check conditional ignorability
- You need to elicit information from domain experts, maybe use the DAG machinery and deduce based on domain knowledge that you have conditional ignorability

If we have observational data Y, D, X, then we should always be controlling for all observed X to estimate the causal effect.



Answer

- No. You need to check (e.g. based on DAGs) that conditional ignorability holds

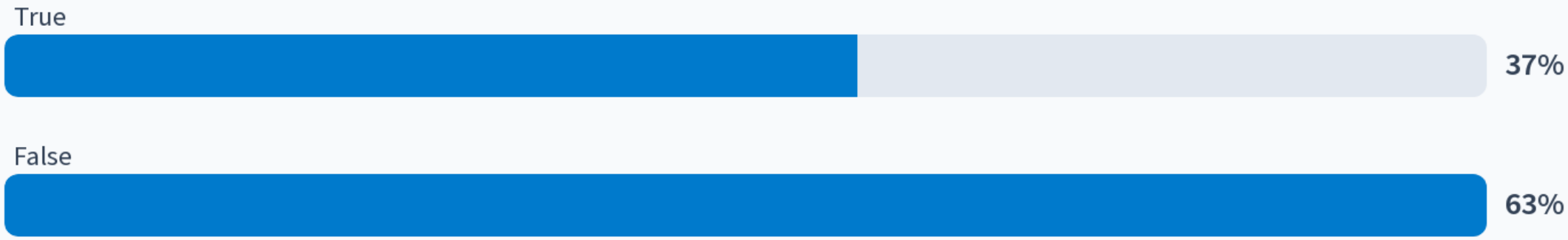
- It’s not always the case that controlling for all X’s is the right thing
- Some X’s can be mediators
- Some X’s can be instruments

We expect conditional ignorability to hold if we condition on a variable that was the result of the treatment.

Answer

- No. Typically conditioning on post-treatment variables is problematic and will lead to bias if our goal is to estimate the ATE (e.g. collider bias)

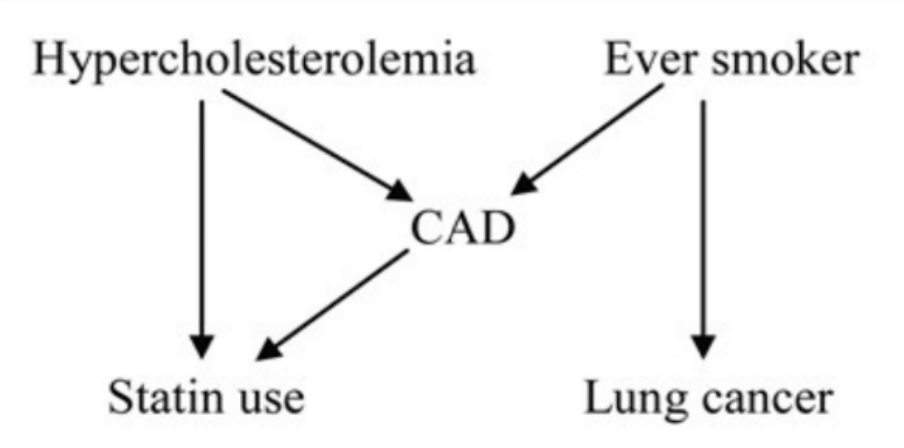
When performing identification by conditioning it is ok to condition on any variable whose value was determined prior to the treatment.



Answer

- Not necessarily; though in practice typically the case
- However, there is the M-bias example, where conditioning on a pre-treatment variable is the wrong thing to do

Assuming the depicted graph is correct and there are no other latent factors; if we estimate the correlation between statin use and lung cancer then this can be interpreted as a causal effect.



Answer

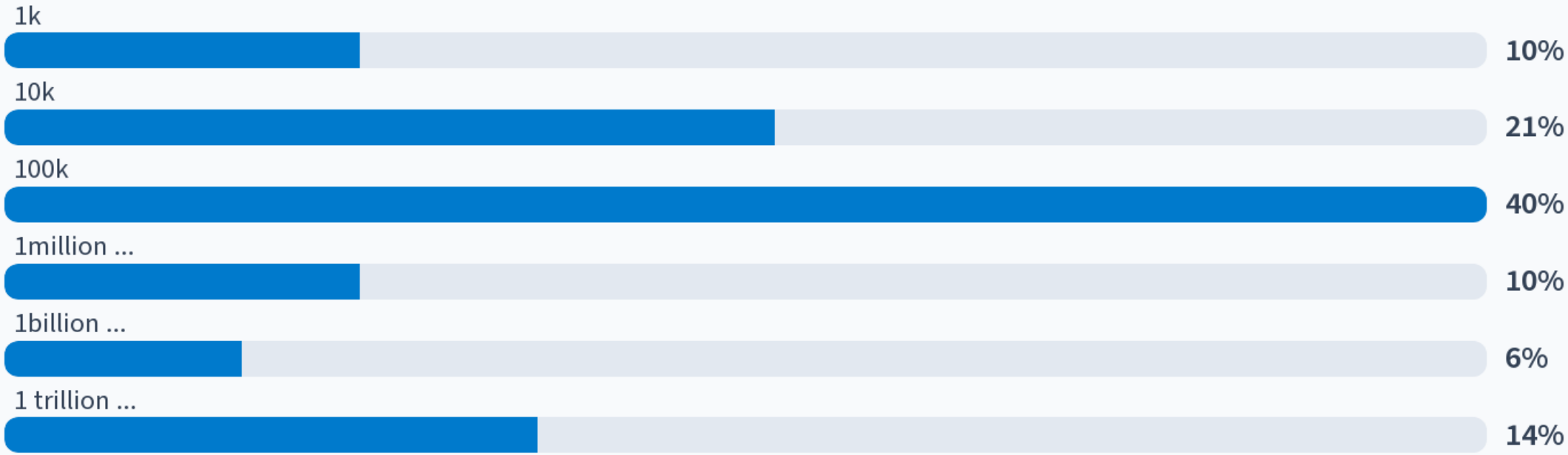
- No. There is a backdoor path through smoking

Suppose that I have conditional ignorability when conditioning on X. Similar to the case of an RCT, without further assumptions, I can estimate the ATE by running OLS of Y using D, X and look at the coefficient of D?

Answer

- The coefficient associated with D, will converge to the ATE only if the true CEF is well-specified, i.e. conditional expectation of Y on X is linear
- Unlike the RCT case (were we did not need well-specification), there is no guarantee otherwise that the coefficient associated with D will converge to the ATE

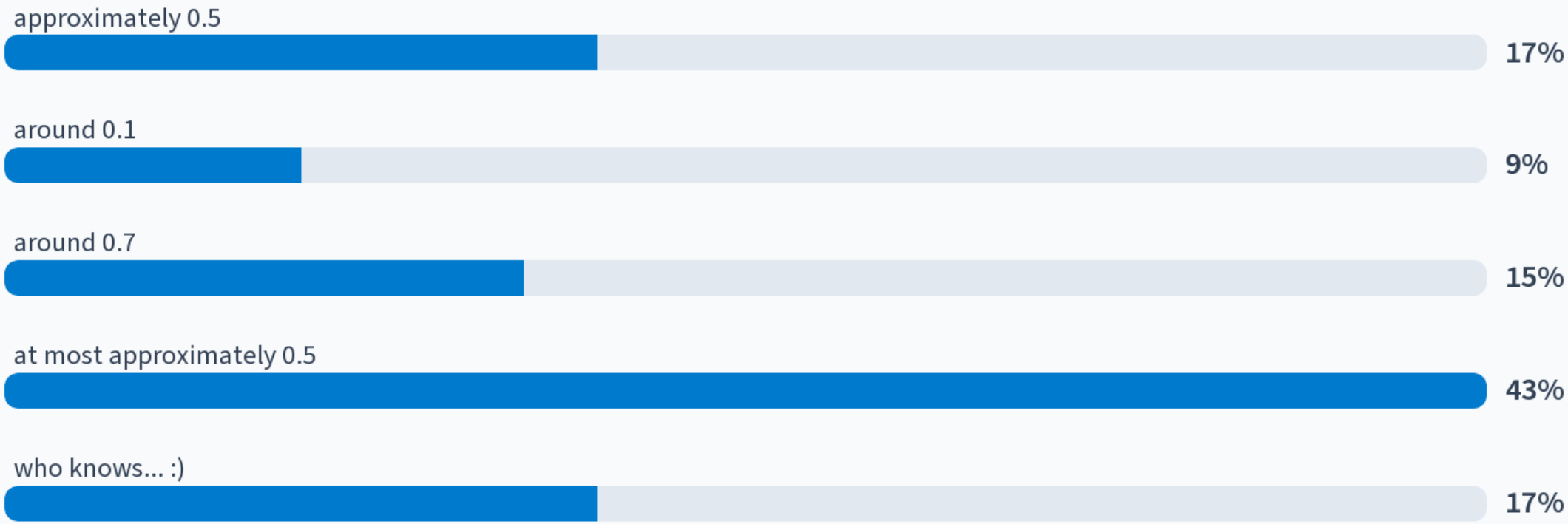
Suppose that I don't just want to estimate the BLP of Y using D, X, but I want to estimate E[Y|D,X]. X contains 10 variables and I'm only willing to assume that E[Y|D,X] is Lipschitz in X. How many samples do I roughly need to guarantee RMSE of 0.1?



Answer

- Based on the non-parametric error rate of $n^{-\frac{1}{p+2}}$ where p is the number of variables you need of the order of a trillion samples…

Suppose that you trained an ML model g for predicting Y from X. You evaluate the model out-of-sample (with 1million samples :) and you get an MSE of 0.5. How far away is the model from the conditional expectation g0(X)=E[Y|X], i.e. E[(g0(X) - g(X))^2]



Answer

- The MSE can be decomposed into the distance and the un-explainable variance

$$E\left[\left(Y - f(X)\right)^2\right] = E\left[\left(f_0(X) - f(X)\right)^2\right] + E[Var(Y|X)]$$

- Since the sum is approximately 0.5, the distance is at most approximately 0.5
- But it can be much smaller. We don’t know how large is the second part

You have two ML models g1, g2. One model has out-of-sample (with 1million samples) MSE of 0.5 and another model of 0.3. How much closer is g2 (than g1) to the conditional expectation function g0, in terms of E[(g0(X) - g(X))^2]?

Answer

- The MSE can be decomposed into the distance and the un-explainable variance

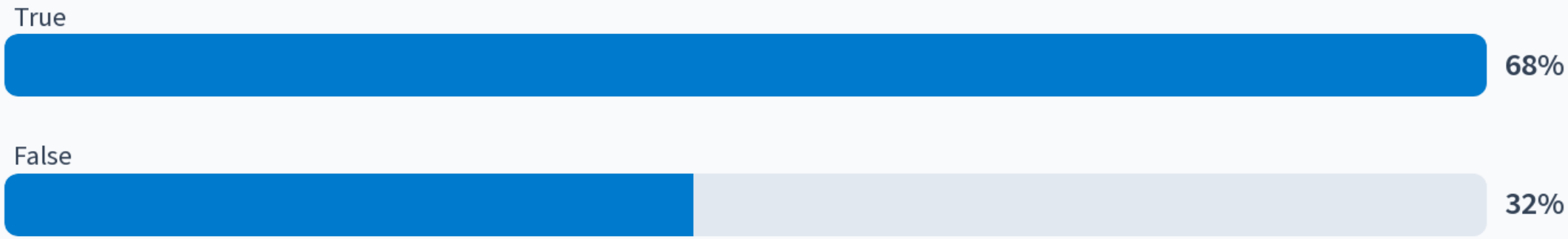
$$E\left[\left(Y - f(X)\right)^2\right] = E\left[\left(f_0(X) - f(X)\right)^2\right] + E[Var(Y|X)]$$

- When we take the difference of the MSE of two models, then the second part cancels

$$E\left[\left(Y - f_1(X)\right)^2\right] - E\left[\left(Y - f_2(X)\right)^2\right]$$
$$= E\left[\left(f_0(X) - f_1(X)\right)^2\right] - E\left[\left(f_0(X) - f_2(X)\right)^2\right]$$

- So the difference in distances is the same as the difference in MSEs

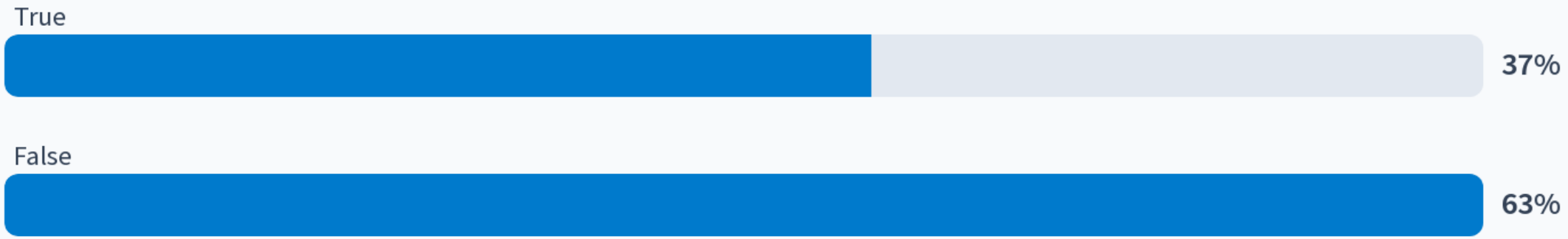
Suppose that a variable D is irrelevant for the prediction of an outcome Y. I regress Y on D, X, where X are other relevant variables, using a Random Forest. I should expect that the distribution of Average Predictive Effect of D will be centered around 0



Answer

- Random Forests introduce bias and can potentially pick up random features as seemingly relevant features.
- Hence, it is highly probable that the effect you will get will be consistently biased and non-zero (unless you perform some form of debiasing; e.g. debiased ML).

In the double lasso algorithm, I can replace the lasso estimator with any generic ML estimator (even automl) and things should work well, without any modification



Answer

- Almost yes..
- The only thing I need to add is sample-splitting, i.e. estimate the automl/forest etc on one sample and run the final OLS on another sample (or do it with cross-fitting)
- Then assuming that the ML methods have fast enough rates, then the guarantees will be the same as in the double lasso

If you have unobserved confounding, you should use the double machine learning methods (e.g. Double ML and Doubly Robust), since the use of ML algorithms, can improve the accuracy of your estimate, as compared to just running OLS.

Answer

- No. Double/debiased ML can only improve your estimation
- It cannot fix your identification
- If you have unobserved confounding, ML will not fix it for you

Assume conditional exogeneity holds when conditioning on X. Suppose I run OLS with interactions, Y = (a + b'X) D + c'X + c0 + eps. Then the function theta(x)=a + b'x is the CATE without any further assumption.

Answer

- No. this is guaranteed to be the case only if the model is well specified, i.e. E[Y|D, X] takes this interactive form

Assume conditional exogeneity holds when conditioning on X. Suppose I run OLS with interactions, Y = (a + b'X) D + c'X + c0 + eps. Then the function theta(x)=a + b'x is the best linear approximation of the CATE without any further assumption.

Answer

- No. At best this is estimating a weighted best linear projection of the CATE, weighted by the variance of the treatment, and even that requires some well-specification assumptions.
- If you want the best linear projection of the CATE then run OLS using the doubly robust labels as target outcomes.

Estimating a heterogeneous treatment effect is equally hard to solving a regression problem.

Answer

- No. It is harder, since we don’t even observe our ideal label which is Y(1) – Y(0).
- Typically involves estimating nuisance/auxiliary models.

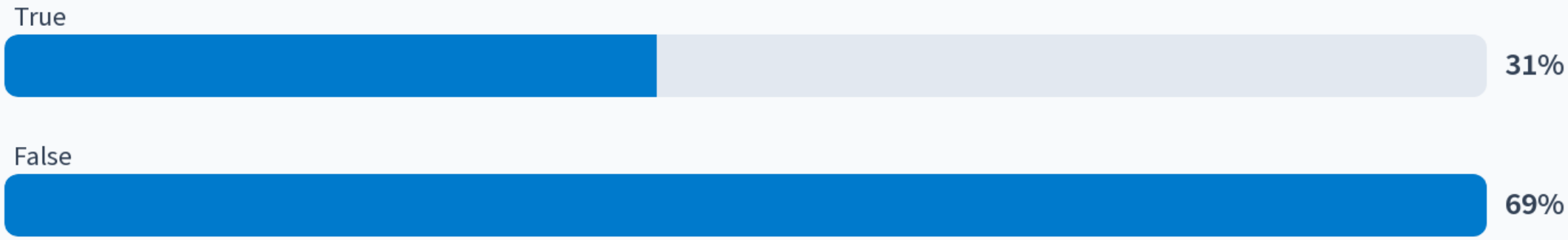
Suppose that my data contain multiple observations from the same unit over a period of time. Then it is easier to perform causal inference.

Answer

- Sometimes yes (e.g. I can check if I can invoke diff-in-diff identification and hence allow for some unobserved confounding)

- Sometimes no (e.g. I can maybe have units undergo multiple treatments in an adaptive manner, in which case I need to invoke more complex dynamic treatment regime identification arguments)

If I have more than two period observations per unit, then the diff-in-siff method can handle any form of unobserved confounding



Answer

- No. The unobserved confounder essentially needs to be persistent and constant for each unit across periods and not time-varying, with the time variation being correlated with D.