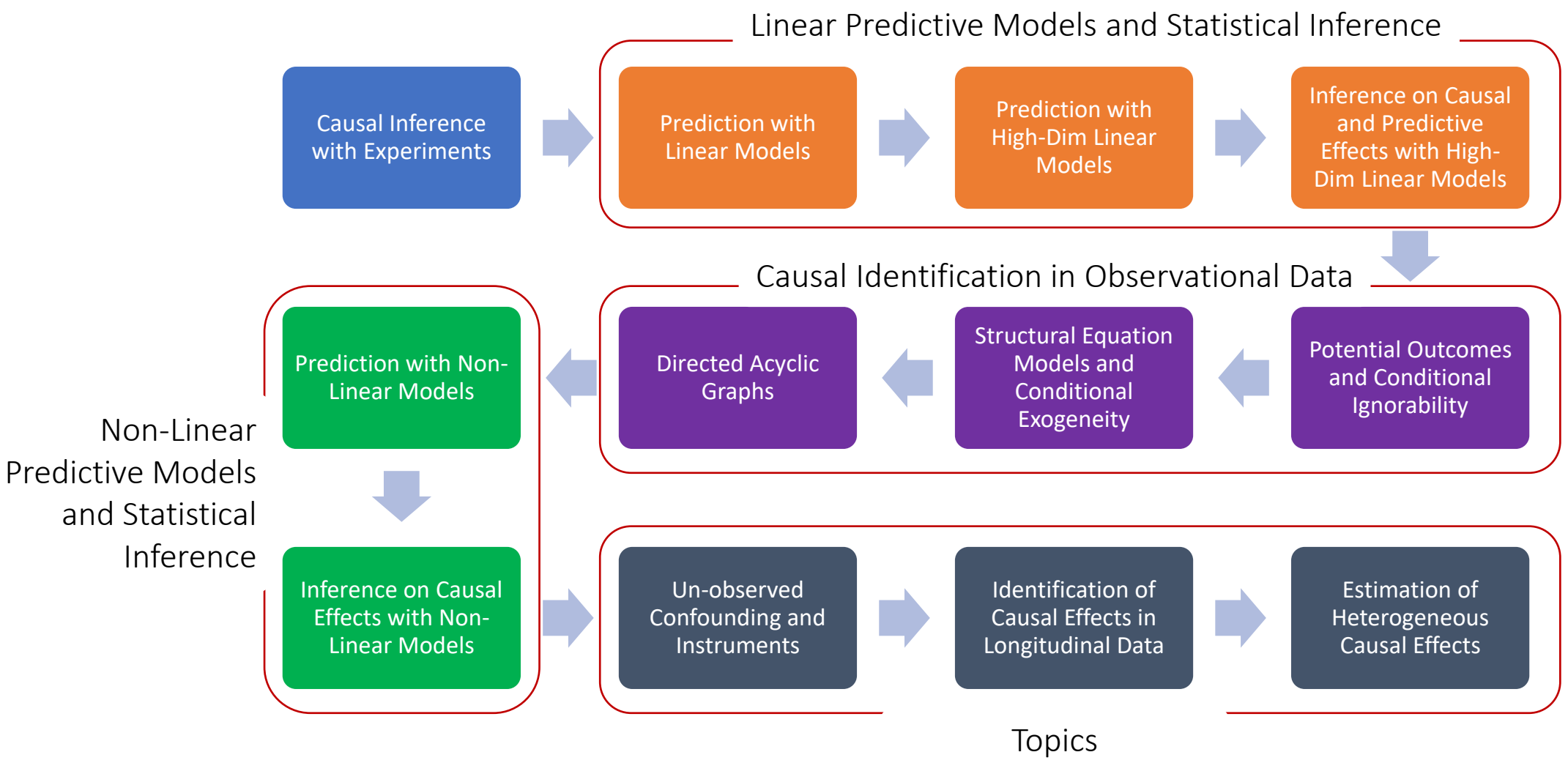
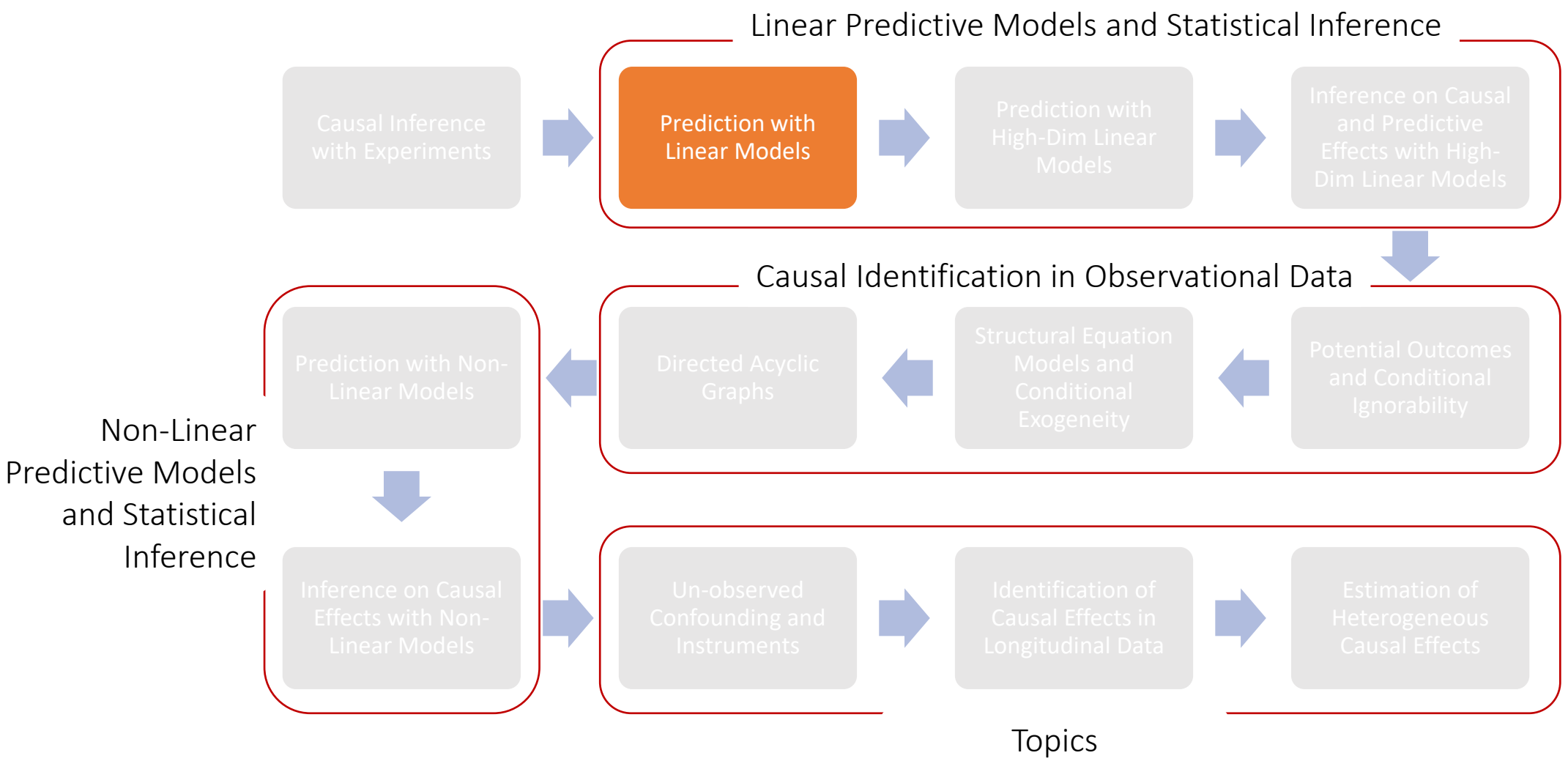
MS&E 228: Analyzing Experiments with Linear Regression

Vasilis Syrgkanis

MS&E, Stanford


Linear Predictive Models and Statistical Inference
Causal Inference with Experiments
Prediction with Linear Models
Prediction with High-Dim Linear Models
Inference on Causal and Predictive Effects with High-Dim Linear Models
Causal Identification in Observational Data
Potential Outcomes and Conditional Ignorability
Structural Equation Models and Conditional Exogeneity
Directed Acyclic Graphs
Non-Linear Predictive Models and Statistical Inference
Prediction with Non-Linear Models
Inference on Causal Effects with Non-Linear Models
Un-observed Confounding and Instruments
Identification of Causal Effects in Longitudinal Data
Estimation of Heterogeneous Causal Effects
Topics


Linear Predictive Models and Statistical Inference
Causal Inference with Experiments
Prediction with Linear Models
Prediction with High-Dim Linear Models
Inference on Causal and Predictive Effects with High-Dim Linear Models
Causal Identification in Observational Data
Potential Outcomes and Conditional Ignorability
Structural Equation Models and Conditional Exogeneity
Directed Acyclic Graphs
Prediction with Non-Linear Models
Non-Linear Predictive Models and Statistical Inference
Inference on Causal Effects with Non-Linear Models
Un-observed Confounding and Instruments
Identification of Causal Effects in Longitudinal Data
Estimation of Heterogeneous Causal Effects
Topics

Even if the relationship between outcome Y and covariates X is non-linear, we can always write:

$$Y = \beta' X + \epsilon, \qquad E[\epsilon X] = 0$$

The function $\beta' X$ is the Best Linear Predictor (BLP) or equivalently the best linear approximation to the Conditional Expectation Function (CEF) $E[Y|X]$

Even if the relationship between outcome Y and covariates X is non-linear, we can always write:

$$Y_i = \hat{\beta}' X_i + \hat{\epsilon}_i, \qquad E_n[\hat{\epsilon} X] = 0$$

The function $\hat{\beta}' X$ is the Best Linear Predictor in sample and $\hat{\beta}$ are the sample regression coefficients

You should expect OLS to produce accurate predictions in the worst-case if the number of variables is small compared to number of samples.

Its predictions converge to the predictions of the BLP in the population

Almost always measure predictive performance of your estimated model on a held-out sample

Predictive effect β_1 of *target variable* is the coefficient in a *simple one variable regression*

$$\begin{pmatrix}\text{part of outcome}\\ \text{un-explained by other}\end{pmatrix} \sim \begin{pmatrix}\text{part of target}\\ \text{un-explained by other}\end{pmatrix}$$

Coefficient of D in OLS(y~D,W) is mathematically equivalent in samples to

yres = y - OLS(y~W).predict(W)
Dres = D - OLS(D~W).predict(W)

Coefficient of Dres in OLS(yres~Dres)

If we want an interval that roughly contains the predictive effect with probability α, we can use

$$CI(\alpha) \coloneqq \left[\hat{\beta}_1 - z_{1-\frac{\alpha}{2}}\hat{\sigma}_n, \hat{\beta}_1 + z_{1-\frac{\alpha}{2}}\hat{\sigma}_n\right]$$

$$\hat{\sigma}_n \coloneqq \frac{1}{\sqrt{n}}\sqrt{\frac{E_n\left[\hat{\epsilon}^2\breve{D}^2\right]}{E_n\left[\breve{D}^2\right]^2}}$$

e.g. for 95% confidence interval, $z_{1-\frac{a}{2}} \approx 1.96$

Revisit Covariate Adjustment for Effect Inference in Experiments

Co-variates for Precision

- Even if we are only interested on ATE covariates can be valuable for precision
- Suppose variance of y is large but can be explained largely by W
- Then we can use W to remove all the explained variation from y
- Then perform our ATE analysis on the remnant variation
- This is oftentimes performed in practice via ordinary linear regression of y on the vector $(1, D, W)$ (after centering W, i.e. $E[W] = 0$)

Is this consistent?

- Suppose that the conditional expectation function (CEF) of the outcome is indeed linear, with $(D, 1, W)$

$$E[\, Y \mid D, W \,] = D\alpha + \alpha_0 + W'\beta$$

- Then note that

$$E[Y(0)] = E\big[E[Y|D = 0, W]\big] = \alpha_0$$
$$E[Y(1)] = E\big[E[Y|D = 1, W]\big] = \alpha + \alpha_0$$

- Baseline outcome is coefficient associated with the intercept 1
- Average effect is coefficient associated with treatment D

Is it consistent if E[Y|D, W] is not linear?

Simple Case

- Suppose we run linear regression of $Y \sim (D, 1)$
- Is the coefficient associated with D, converging to the ATE?
- Let's look at the population BLP:

$$Y = a_0 + aD + \epsilon, \qquad E[\epsilon\ (D, 1)] = 0$$

- Let's examine the normal equations:

$$E[\epsilon\ D] = 0 \Rightarrow E[\epsilon \mid D = 1] = 0 \Rightarrow E[\,Y \mid D = 1\,] = a_0 + a$$
$$E[\epsilon] = 0 \Rightarrow E[\,\epsilon \mid D = 0\,] = 0 \Rightarrow E[\,Y \mid D = 0\,] = a_0$$

- From these two we uncover that:

$$a = E[\,Y \mid D = 1\,] - E[\,Y \mid D = 0\,] = E[Y(1) - Y(0)] = \delta$$

What if we add covariates?

- By the BLP decomposition of Y using $(D, 1, W)$

$$Y = D\alpha + \alpha_0 + \beta' W + \epsilon, \qquad E[\epsilon\ (D; 1; W)] = 0$$

- Note that the quantity:

$$U = \beta' W + \epsilon$$

- Also satisfies

$$E[U\ (D; 1)] = \beta' E[W\ (D; 1)] + E[\epsilon(D; 1)]$$
$$= \beta' E[W]\ E[(D; 1)] = 0$$

$= 0$ by orthogonality of ϵ

By independence of W and D

Since $E[W] = 0$ (de-meaned W)

Is this consistent? Beyond Linear CEF

- We can write

$$Y = D\alpha + \alpha_0 + U, \qquad E[U\ (D; 1)] = 0$$

- where $U = \beta' W + \epsilon$
- Thus α, α_0 solve the Normal Equations of Y on $D, 1$
- Thus α, α_0 are the BLP of Y using $(D, 1)$

- The parameters a, a_0 are the same as the parameters in the population BLP if we did not use covariates!

- We already saw that in that case a is the ATE!

Is this consistent? Beyond Linear CEF

- The crucial property we used was that $U = \beta' W + \epsilon$ is uncorrelated with $(D, 1)$
- This holds whenever W is un-correlated with $(D, 1)$

- So, if we run OLS with any set of controls W that are un-correlated with $(D, 1)$, then the population parameters a, a_0 associated with $(D, 1)$, remain un-changed!

The coefficient associated with treatment D in OLS with co-variate adjustment is always consistent for the treatment effect, when run on data from a randomized experiment, as-long-as covariates are de-meaned. The true relationship of outcome with covariates does not need to be linear.

Analysis of Variance (ANOVA)

Quality of In-Sample Estimate

- What if we run a Linear Regression with n samples
- How accurate is the parameter $\hat{a}$ in the in-sample BLP?
- Does the accuracy change or improve if we use W or not?

Variance of Estimate

- The parameter a is the "predictive effect" of D
- Using the characterization of the estimate of the predictive effect:

$$\sqrt{n}(\hat{\alpha} - \alpha) \overset{\text{a}}{\sim} N(0, V_\alpha) \qquad V_\alpha = \frac{E\left[\epsilon^2 \widetilde{D}^2\right]}{E\left[\widetilde{D}^2\right]^2}$$

- ϵ is population residual outcome:

$$\epsilon = y - D\alpha - \alpha_0 - W'\beta \qquad E[\epsilon(D, 1, W)] = 0$$

- $\widetilde{D}$ is residual treatment (removing whatever is linearly predictable from $(1, W)$)

$$\widetilde{D} = D - E[D]$$

Heteroskedasticity Robust Variance

- Variance formula

$$V_\alpha = \frac{E\left[\epsilon^2 \widetilde{D}^2\right]}{E\left[\widetilde{D}^2\right]^2}$$

- is valid even when the linear CEF assumption is violated
- Important to note that this formula is known as the "heteroskedasticity robust variance formula" (HC0)
- Many software packages make the simplification that the residual ϵ is independent of D, W, leading to $V_\alpha = E\left[\epsilon^2\right]/E\left[\widetilde{D}^2\right]$. This is incorrect in most cases!

Variance without adjustment

- If we don't use W, then the formula remains the same, but the residual changes (note: population parameters a, a_0 are the same)

$$\bar{\epsilon} = Y - D\alpha - \alpha_0 = W'\beta + \epsilon$$

- The residual treatment $\widetilde{D}$ is the same, since W is independent of D
- Comparing the two variances, only the numerators change:

$$E\left[\bar{\epsilon}^2\widetilde{D}^2\right] = E\left[(W'\beta + \epsilon)^2\widetilde{D}^2\right]$$
$$= E\left[(W'\beta)^2\widetilde{D}^2\right] + E\left[\epsilon^2\widetilde{D}^2\right] + 2E\left[\beta'W\epsilon\widetilde{D}^2\right]$$
$$= E\left[(W'\beta)^2\widetilde{D}^2\right] + E\left[\epsilon^2\widetilde{D}^2\right]$$

Numerator when not using W

Numerator when using W

If this term is zero, then we know that using W can only decrease variance of estimate

When can we claim that interaction term is zero?

$$E\left[\epsilon W \tilde{D}^2\right] = 0$$

One Case: Strong Assumption

- Suppose that CEF is linear, i.e.

$$E[\,Y \mid D, W\,] = a_0 + a\,D + \beta' W$$

- Then we have that:

$$Y = a_0 + a\,D + \beta' W + \epsilon, \qquad E[\,\epsilon \mid D, W\,] = 0$$

- Then we can verify that the cross-term is zero:

$$E\left[\epsilon W \widetilde{D}^2\right] = E\left[E[\,\epsilon \mid D, W\,] W \widetilde{D}^2\right] = 0$$

- Hence, when CEF is linear then variance of OLS estimate with extra co-variates (adjusted) is weakly smaller than two-means estimate (un-adjusted)

What if CEF is Non-Linear? Example: Heterogeneous Effects

- Assume $E[W] = 0$ and suppose that:

effect modifier

$$E[\,Y \mid D, X\,] = D\alpha + \alpha_0 + DW'\gamma + W'\beta$$

ATE

- Equivalently:

$$Y = Da + a_0 + DW'\gamma + W'\beta + \nu, \qquad E[\,\nu \mid D, W\,] = 0$$

- What OLS $Y \sim D, 1, W$ is estimating is coefficients $\tilde{\alpha}, \tilde{\alpha}_0, \tilde{\beta}$, the solution to:

$$E\left[\left(Y - D\tilde{\alpha} - \tilde{\alpha}_0 - W'\tilde{\beta}\right)\begin{pmatrix} D \\ 1 \\ W \end{pmatrix}\right] = 0$$

- Since W is un-correlated with $D, 1$ by our previous analysis, $\tilde{a}, \tilde{a}_0$ are the same as a, a_0, i.e. we are still recovering the correct ATE.

What if CEF is Non-Linear? Example: Heterogeneous Effects

- Assume $E[W] = 0$ and suppose that:

$$Y = Da + a_0 + DW'\gamma + W'\beta + \nu, \qquad E[\,\nu \mid D, W\,] = 0$$

- What OLS $Y \sim D, 1, W$ is estimating is coefficients $\tilde{\alpha}, \tilde{\alpha}_0, \tilde{\beta}$, the solution to:

$$E\left[\left(Y - D\tilde{\alpha} - \tilde{\alpha}_0 - W'\tilde{\beta}\right)\begin{pmatrix} D \\ 1 \\ W \end{pmatrix}\right] = 0$$

- But $\tilde{\beta} = \beta + E[D]\gamma$ by examining the third normal equation:

$$E\left[\left(DW'\gamma + W'\beta\ - W'\tilde{\beta}\right)W\right] = 0 \Rightarrow E[W']\left(\gamma E[D] + \beta - \tilde{\beta}\right) = 0$$

- So, the residual of OLS of $Y \sim D, 1, W$ is:

$$\epsilon = Y - Da - a_0 - W'\tilde{\beta} = (D - E[D])W'\gamma + \nu, \qquad E[\,\nu \mid D, W\,] = 0$$

- And the cross-term in the variance analysis becomes:

$$\mathrm{E}\left[W\epsilon\widetilde{D}^2\right] = E\left[\widetilde{D}^3\right]E[WW']\gamma \neq 0$$

OLS with Interactive Terms

- What if instead we run OLS including interaction terms [Lin'13]

$$Y \sim D, 1, W, DW$$

- Since $E[W] = 0$ and $W \perp\!\!\!\perp D$, we have (DW, W) are un-correlated with D

$$E[DDW] = E\left[D^2\right]E[W] = 0$$

- Thus, by the analysis we did before, in the absence of any model assumptions, the coefficient of D and of the intercept, recover the true ATE a and the true mean baseline outcome a_0

Let's try it out!

OLS with Interactive Terms

- These interactive terms enforce the residual ϵ of OLS to satisfy a stronger orthogonality property

$$E\left[\epsilon\begin{pmatrix} D \\ 1 \\ W \\ DW \end{pmatrix}\right] = 0$$

- We can conclude the stronger un-correlatedness properties:

$$E[\epsilon DW] = 0 \Rightarrow E[\,\epsilon W \mid D = 1\,] = 0$$
$$E[\epsilon W] = 0 \Rightarrow E[\,\epsilon W \mid D = 0\,] = 0$$

- From which we can argue that the cross-term in the variance is zero

$$\mathrm{E}\left[W\epsilon\widetilde{D}^2\right] = E\left[E[\,W\epsilon \mid D\,]\widetilde{D}^2\right] = 0$$

Even if you only care about ATE, if you have p covariates and $p \ll n$ run OLS with interactive terms (after de-meaning covariates)!

Guaranteed improved precision