Linear Regression for Causal Analysis

$$\boxed{Y = \alpha T + \beta X + \varepsilon}$$

predictors

outcome / response var — treatment — covariates

ASSUMPTION: All independent variables are uncorrelated with error term.

Endogeneity

- Predictor variable correlated with error term → endogeneity
- Coefficient of this predictor no longer BLUE (Best Linear Unbiased Estimator) → violates assumption of linear regression

Var not determined by any other variables in eqn. → exogeneity

Sources of Endogeneity

Unmeasured confounders

- Unmeasured confounder → treatment
  → outcome
- If confounder emitted → leaks into error term → predictor corr. w/ error → endogeneity
- Capture all confounders to remove

Measurement errors

- All vars should have accurate values
- Sometimes proxy vars used → does not reflect true value of qty to be measured → Error = true - observed

Error in dependent variable Y ✓
↓
does not depend on explanatory variables
↓
exogeneous

Error in predictor variables ✗
↓
Relationship present but not captured
↓
Unmeasured confounder

Simultaneity

Y ⇄ X

If X determined by Y → $\underbrace{\text{X corr. w/ error term}}_{\text{endogeneity}}$ } Keep Y fixed and try changing X, see what happens

Controlling Endogeneity

- Simply having all confounders in eqn not sufficient
  - unmeasured confounders
  - simultaneity
  - measurement errors
- Introduce one or more instrumental vars using Two Stage Least Squares

OLS: Recap

$$\boxed{y = \beta X + \varepsilon}$$

$$\min_{\beta} (y - \beta X)^T (y - \beta X) \quad \Big\} \text{ minimise } (y - \hat{y})^2$$

Solution: $\beta^* = \underbrace{(X^T X)^{-1} X^T}_{\text{Moore-Penrose pseudoinverse}} y$

Frisch - Waugh - Lovell Theorem

Context

Simple bivariate case w/ treatment and outcome:

$$\boxed{y_i = \beta_0 + \beta_1 T_i + \varepsilon}$$

If treatment truly random → independent of ε → exogeneous

$$\Rightarrow \text{ATE} = E[Y(1) - Y(0)] = E[(\beta_0 + \beta_1) - (\beta_0 + \beta_1(0))] = E[(\beta_0 + \beta_1) - (\beta_0)] = E[\beta_1] = \beta_1$$

However, this is not usually the case.

We will have additional covariates that we need to condition for to simulate independence.

conditional independence assumption → not directly testable

Theorem

Estimate any key parameter of linear regression by first "partialling out" the effects of additional covariates

Procedure

Say you want to analyse the effect of treatment T on outcome Y.

① Regress outcome and treatment on covariate(s) X

$$\hat{Y} = \alpha_0 + \alpha_1 X + \varepsilon$$
$$\hat{T} = \gamma_0 + \gamma_1 X + \eta$$

② Get residuals

$$Y_i^* = Y_i - \hat{Y}_i \quad \Big\} \text{ variation in Y not explained by X}$$
$$T_i^* = T_i - \hat{T}_i \quad \Big\} \text{ variation in T not explained by X}$$

③ Regress Y^* on T^*

$$Y_i^* = \beta_0 + \beta_1 T_i^* + \varepsilon^* \quad \Big\} \text{ only includes portions of Y and T that are independent from covariates}$$

Debiasing: T on X
Denoising: Y on X
Outcome model: Y^* on T^*

Note:

By FWL, the following estimators of β_1 are equivalent:

- Y on T and X
- Y on T^*
- Y^* on T^*

Double ML

Same thing as FWL but ML models instead of OLS

$$\boxed{Y - M_y(X) = \beta_0 + \beta_1 [T - M_T(X)] + \varepsilon}$$

Doubly Robust (DR) Methods

- Provide consistent estimates if at least 1 out of 2 models correctly specified.
- Estimate two "nuisance" models
- DML = DR with ML for "nuisance" models
  either outcome / treatment model correct → consistent estimates
- Outcome model: Y on T, X
- Treatment model / propensity score model: W_i / T on X