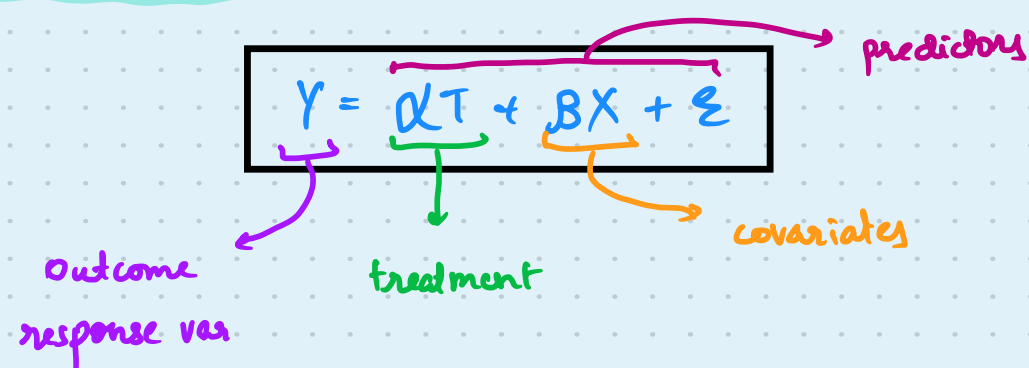


# LINEAR REGRESSION FOR CAUSAL ANALYSIS



ASSUMPTION: All independent variables are uncorrelated with error term.

## Endogeneity

- Predictor variable correlated with error term  $\rightarrow$  endogeneity
- Coefficient of this predictor no longer BLUE (Best Linear Unbiased Estimator)  $\rightarrow$  violates assumption of linear regression
- Var not determined by any other variables in eqn.  $\rightarrow$  exogeneity

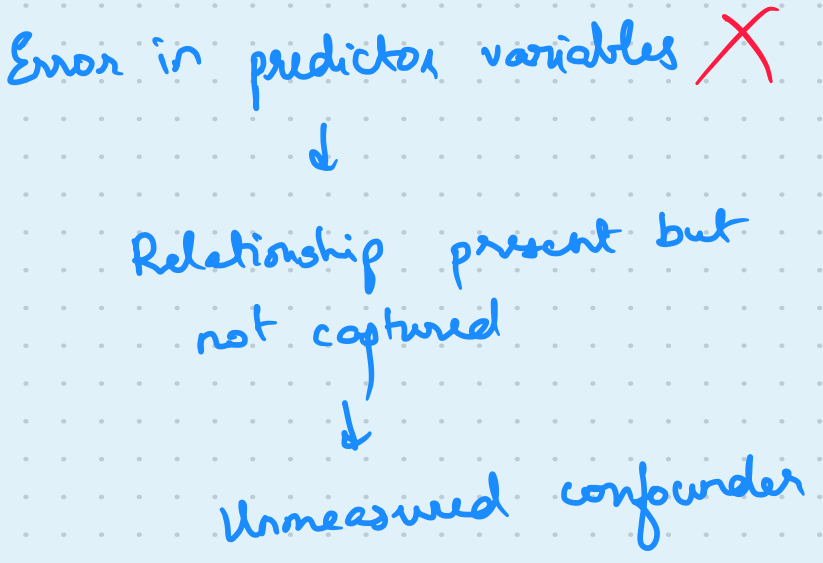
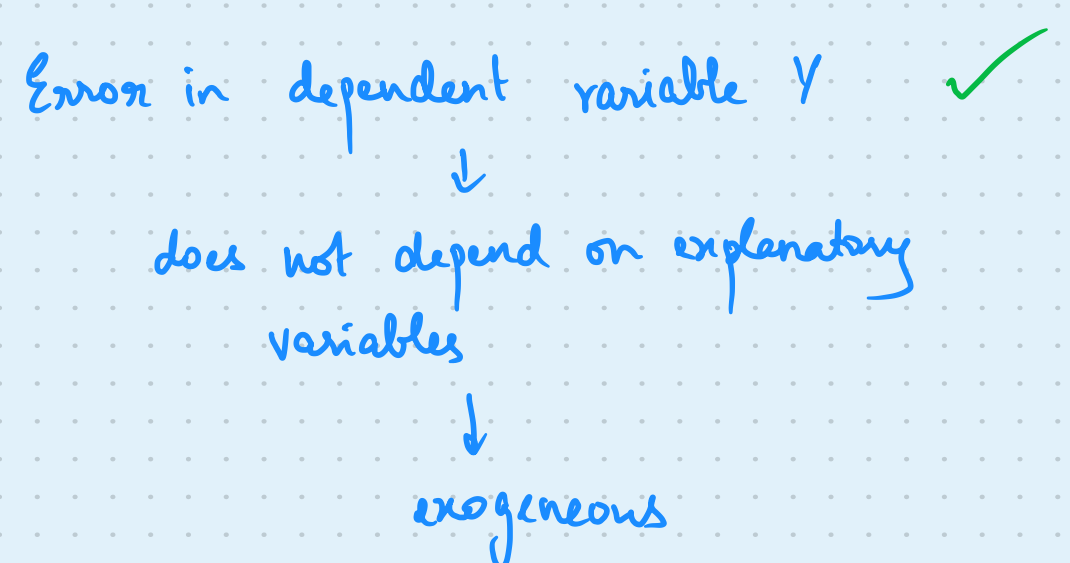
## Sources of Endogeneity

### Unmeasured confounders

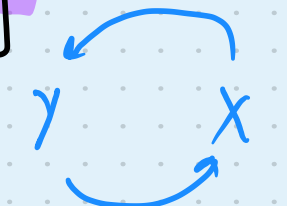
- Unmeasured confounder  $\rightarrow$  treatment  $\rightarrow$  outcome
- If confounder omitted  $\rightarrow$  leaks into error term  $\rightarrow$  predictor corr. w/ error  $\rightarrow$  endogeneity
- Capture all confounders to remove

### Measurement error

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



## Simultaneity



If X determined by Y  $\rightarrow$  X corr. w/ error term  $\rightarrow$  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 error
- Introduce one or more instrumental vars using Two Stage Least Squares

## OLS: Recap

$$y = \beta X + \epsilon$$

$$\min_{\beta} (y - \beta X)^T (y - \beta X) \quad \left\{ \begin{array}{l} \text{minimise } (y - \hat{y})^2 \end{array} \right.$$

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

## FRISCH - WAUGH - LOVELL THEOREM

### Context

Simple bivariate case w/ treatment and outcome:

$$y_i = \beta_0 + \beta_1 T_i + \epsilon$$

If treatment truly random  $\rightarrow$  independent of  $\epsilon \rightarrow$  exogenous

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

However, this is not usually the case.

We will have additional covariates that we need to condition for to simulate independence.  
conditional independence assumption  $\rightarrow$  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  
 $\hat{Y} = \alpha_0 + \alpha_1 X + \epsilon$   
 $\hat{T} = \gamma_0 + \gamma_1 X + \eta$

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

- 2 Get residuals  
 $Y_i^* = Y_i - \hat{Y}_i$  } variation in Y not explained by X  
 $T_i^* = T_i - \hat{T}_i$  } variation in T not explained by X

- 3 Regress  $Y^*$  on  $T^*$   
 $Y_i^* = \beta_0 + \beta_1 T_i^* + \epsilon^*$  } only includes portions of Y and T that are independent from covariates

NOTE:  
By FWL, the following estimators of  $\beta_1$  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

$$Y - M_Y(X) = \beta_0 + \beta_1 [T - M_T(X)] + \epsilon$$

## DOUBLE 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  $\rightarrow$  consistent estimates
- Outcome model: Y on T, X
- Treatment model / propensity score model: W; T on X