Lecture 7: Assessing Regression Studies(I)

Introduction to Econometrics,Spring 2026

Zhaopeng Qu

Business School,Nanjing University

April 23 2026

Review of Previous Lectures

Simple and Multiple OLS Regressions

- The simple OLS regression model is

$$Y_i = \beta_0 + \beta_1 X_i + u_i, i = 1, ..., n$$

- The OLS estimator for β_1 is

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_i - \bar{X})^2}$$

- The OLS regression model is

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + ... + \beta_k X_{k,i} + u_i, i = 1, ..., n$$

- The OLS estimator for β_j

$$\hat{\beta}_j = \frac{\sum_{i=1}^{n} \tilde{X}_{j,i} Y_i}{\sum_{i=1}^{n} \tilde{X}_{j,i}^2} \; for \; j = 1, 2, .., k$$

Simple/Multiple OLS Regressions: Assumptions

- If the four least squares assumptions in the multiple regression model hold:
  - **Assumption 1**: The conditional distribution of u_i given $X_{1i}, ..., X_{ki}$ has mean zero,thus

$$E[u_i|X_{1i}, ..., X_{ki}] = 0$$

  - **Assumption 2**: $(Y_i, X_{1i}, ..., X_{ki})$ are i.i.d.
  - **Assumption 3**: Large outliers are unlikely.
  - **Assumption 4**: No perfect multicollinearity.(only for Multiple OLS regression)
- Then
  - The OLS estimators $\hat{\beta}_0, \hat{\beta}_1 ... \hat{\beta}_k$ are unbiased.
  - The OLS estimators $\hat{\beta}_0, \hat{\beta}_1 ... \hat{\beta}_k$ are consistent.
  - The OLS estimators $\hat{\beta}_0, \hat{\beta}_1 ... \hat{\beta}_k$ are normally distributed in large samples.

Nonlinear Regression Model

1. *Nonlinear in Xs*
    - Polynomials,Logarithms and Interactions
    - The multiple regression framework can be extended to handle regression functions that are nonlinear in one or more X.
    - the difference from a standard multiple OLS regression is *how to explain estimating coefficients*.

Nonlinear Regression Model

2. *Nonlinear in* β or *Nonlinear in Y*
    - Discrete Dependent Variables or Limited Dependent Variables.
    - Linear function in Xs is not a good prediciton function or Y, instead a function which parameters enter nonlinearly, such as logisitic or standard normal.
    - The parameters can not obtained by OLS estimation any more but **Maximum Likelihood Estimation**

- MLE need more assumptions than OLS, thus the distribution of u_i is known.
- In the large sample, the MLE estimator is also consistent and asymptotically normal

Accessing Regression Studies: Introduction

Definitions of Validity

- The concepts of **internal and external validity** provide a general framework for assessing whether a empirical studies answers a specific question of interest rightly and usefully.
  - **Internal validity**:

- **External validity**:

- Internal and external validity distinguish between
  - *the population and setting being studied*
  - *the population and setting to which the results are generalized.*

Differences between studied and interest

- **The population and setting studied**
  - The population studied is the population of entities-people, companies, school districts, and so forth-from which the sample is drawn.
  - The setting studied refers to as the institutional, legal, social, and economic environment in which the population studied fits in and the sample is drawn.
- **The population and setting of interest**
  - The population and setting of interest is the population and setting of entities to which the causal inferences from the study are to be applied(generalized).
- Example: Class size and test score
  - the population studies: elementary schools in CA
  - the population of interest: middle schools in CA
  - different populations and settings: elementary schools in MA

Warp up

- **Internal validity** is the top priority in causal inference studies.
- **External validity** is the secondary focus, but only if internal validity is secured.
- In result, we care about the internal validity over 100 times than the external validity in most studies.
- The following content will focus on the internal validity of regression studies.

Threats to Internal Validity in OLS Regressions

- Suppose we are interested in the causal effect of X_1 on Y and we estimate the following multiple regression model

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \ldots + \beta_k X_{k,i} + u_i, i = 1, \ldots, n$$

- Internal validity has **three components**:

Threats to Internal Validity

- Threats to internal validity:

- In a narrow sense,
  - **Internal Invalidity = endogeneity in the estimation** which is caused by the above 1-5 threats.
- In a broad sense,
  - **Internal Invalidity = 1-5 threats + 6-7 threats**

Omitted Variable Bias(OVB) and Control Variables

OVB Review

- Suppose we want to estimate the causal effect of X_i on Y_i,which represent STR and Test Score,respectively.
- Besides, W_i is the share of English learners which is **omitted** in the regression.
- Two models are as follows:

OVB Review

- Then we have the OLS estimator of β_1 as follows

- An omitted variable W_i leads to an inconsistent OLS estimate of the causal effect of X_i if **both**

- The OLS estimator does not provide a unbiased and consistent estimate of the causal effect of X_i, in other words,the OLS regression is not **internally valid**.

Wrap Up

- OVB bias is the most possible bias when we run OLS regression using nonexperimental data.
- OVB bias means that there are some variables which **should** have been included in the regression but actually was not.
- Then the simplest way to overcome OVB: **Control**
  - Putting omitted variables into the right side of the regression as **control variables**, which are independent variables that are not the variable of interest.
- A critical question, often overlooked by many students and even experienced researchers, is:
  - Should we control as many variables as possible to avoid omitted variable bias (OVB)?
  - What kinds of variables can serve as control variables?
- Let us dig deeper into the **control variables** and **OVB** in the following sections.

OLS Regression Estimators in Partitioned Regression

Recall: OLS Regression Estimators in Multiple OLS

- OLS estimator in Multiple OLS

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + ... + \beta_k X_{i,k} + u_i, i = 1, ..., n$$

- The OLS estimator of β_j is

$$\hat{\beta}_j = \frac{\sum_{i=1}^{n} \tilde{X}_{ij}, Y_i}{\sum_{i=1}^{n} (\tilde{X}_{ij})^2}$$

- The asymptotic OLS estimator of β_j

- Where $\tilde{X}_{ij}$ is the fitted OLS **residual** of regressing X_{ij} on other regressors,thus

Recall: Unbiasedness of OLS Estimators

- Based on the multiple OLS estimators in ,we have

- For simplicity, under the 5th assumption of multiple OLS regression: **homoskedasticity**, thus

Recall: the Standard Error of $\hat{\beta}$

- Then we have the variance of $\hat{\beta}_j$ as follows

Recall: the Standard Error of $\hat{\beta}$ and R^2

- The $\tilde{X}_{ij}$ is the **residual** obtained from the regression of X_j on all other Xs

- Then the **R-Squared** of this partitioned regression is

- Where SSR_j is the sum of the squared residuals and TSS is the total variances of X_j.And $s_j^2 = \frac{\sum_{i=1}^{n}(X_{ji}-\bar{X}_j)^2}{n-1}$ is the sample variance of X_j.

Recall: the Standard Error of $\hat{\beta}$

The Variance of $\hat{\beta}_j$ under Homoskedasticity

Recall: the Standard Error of $\hat{\beta}$

The Variance of $\hat{\beta}_j$ under Homoskedasticity

- How does the variance of $\hat{\beta}_j$ change with the following factors?

Recall: the Standard Error of $\hat{\beta}$

The Variance of $\hat{\beta}_j$ under Homoskedasticity

- How does the variance of $\hat{\beta}_j$ change with the following factors?

Factors	symbols	$Var\left(\hat{\beta}_j\right)$
the variance of u_i		
the sample variance of X_j		
the R_j^2		
the sample size		

Control Variables

Control Variables: W

- The basic regression model is

$$Y_i = \beta_0 + \beta_1 X_i + \gamma W_i + u_i$$

  - X_i is the variable of interest and W_i is the **control variable** which is **NOT** the variable of interest.
- Based on the relationships between W and Y and W and X, we can classify the control variables into several categories:
  1. Whether W has an effect on Y or not:
     - **Relevant Variables**:
     - **Irrelevant Variables**:
  2. Whether W is correlated with X or not:
     - **Uncorrelated Control Variables**:
     - **Correlated Control Variables**:
     - **Highly-Correlated Control Variables**:

Control Variables: W

- We will discuss the potential bias or the precise of the OLS estimators when the control variable is as follows:

Control Variables: Class size on Test scores

- Tell me following variables can be classified into which categories?
    1. gender composition of the class
    2. the weather(temperature) of the exams
    3. the share of English learners in the class
    4. the size of the classroom

Irrelevant Variables: Models

- Assume that we have an **irrelevant** control variable W into the model, thus the model is

$$Y_i = \beta_0 + \beta_1 X_i + \gamma W_i + u_i \tag{7.1}$$

- Since W is **irrelevant** to Y, thus

- Then the model excluding W is

- where $v_i = \gamma W_i + u_i$, then $\beta_0 = \tilde{\beta}_0$ and $\beta_1 = \tilde{\beta}_1$.

Irrelevant Variables: Estimate

- Then based on the OVB formula for (7.2),we have

- The OLS estimator $\hat{\beta}_1$ is still **consistent** whether you include W or not.

Irrelevant Variables: Variance

- The variance of $\hat{\beta}_1$ in 7.1 is

Where R^2_{xw} is the R-Squared of the regression of X on W

- The variance of $\hat{\hat{\beta}}$ in 7.2 is

Irrelevant Variables: Variance

- **Because** $v_i = \gamma W_i + u_i$ **and** $\gamma = 0$, **thus**

- **Based on the relationship between** W **and** X **in three cases as follows**

$$Cov(X_i, W_i) \begin{cases} = 0 & \text{if } X_i \text{ is not correlated with } W_i \\ \neq 0 & \text{if } X_i \text{ is correlated with } X_i \\ \to 1 & \text{if } X_i \text{ is highly correlated with } X_i : \text{multicollinearity} \end{cases}$$

- **Then**

$$R^2_{xw} \begin{cases} = 0 & \text{if } X_i \text{ is not correlated with } W_i \\ \neq 0 & \text{if } X_i \text{ is correlated with } X_i \\ \neq 0 \text{ } and \text{ } \rho \to 1 & \text{if } X_i \text{ is highly correlated with } X_i : \text{multicollinearity} \end{cases}$$

Highly-Correlated Variables

- Recall: **Perfect multicollinearity** arises when one of the regressors is a **perfect** linear combination of the other regressors.
- If you include a full set of binary variables (a complete and mutually exclusive categorization) and an intercept in the regression, you will have *perfect multicollinearity*.

Multicollinearity

Multicollinearity means that two or more regressors are **highly correlated**, but one regressor is **NOT** a perfect linear function of one or more of the other regressors(NOT perfect multicollinearity).

- **Multicollinearity** is **NOT** a violation of OLS assumptions.
  - It does not impose theoretical problem for the calculation of OLS estimators.
- But if two regressors are highly correlated, then the the coefficient on at least one of the regressors is **imprecisely estimated (high variance)**.
- Recall: the variance of $\hat{\beta}_j$ is

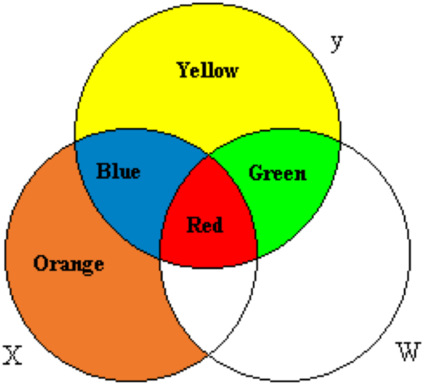
  - When R_j^2 is high, the variance of $\hat{\beta}_j$ is high.

Multicollinearity

- To what extent two correlated variables can be seen as “highly correlated”?
- **Rule of Thumb**:
  - **Mild multicollinearity**: correlation of 0.5-0.7
  - **High multicollinearity**: correlation above 0.7-0.8
  - **Severe multicollinearity**: correlation above 0.9
- **Variance Inflation Factor (VIF)** is a more comprehensive measure:

- where R_j^2 is the R-Squared of the regression of X_j on all other Xs.
- **Rule of Thumb**:
  - **Generally acceptable**: $VIF < 5$
  - **Moderate multicollinearity**: $5 \leq VIF < 10$
  - **Severe multicollinearity**: $VIF \geq 10$

Venn Diagrams for Multiple Regression Model



- In a simple model (y on X), OLS uses ‘Blue‘ + ‘Red‘ to estimate β.
- When y is regressed on X and W: OLS throws away the red area and just uses blue to estimate β.
- Idea: Red area is contaminated(we do not know if the movements in y are due to X or to W).

Venn Diagrams for Multicollinearity

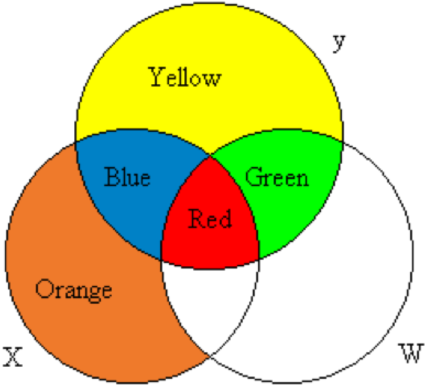


Figure 3a Modest collinearity

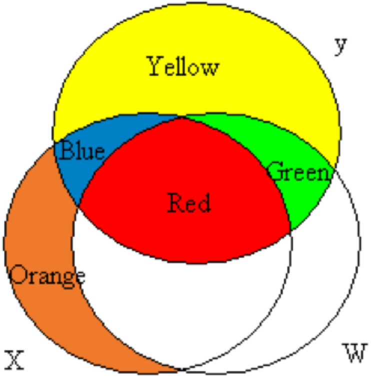


Figure 3b Considerable collinearity

Venn Diagrams for Multicollinearity

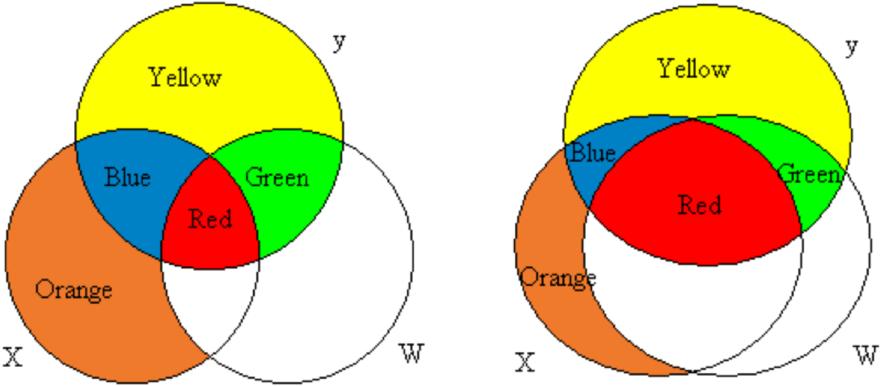


Figure 3a Modest collinearity **Figure 3b Considerable collinearity**

- Less information (compare the Blue and Green areas in both figures) is used, the estimation is less precise.

Irrelevant Variables: Warp Up

- Then we have
  1. if irrelevant variable W_i is **not** correlated with X_i, then $R^2_{xw} = 0$ and

2. if irrelevant variable W_i is **correlated** with X_i, then $R^2_{xw} \neq 0$ and

3. if irrelevant variable W_i is **highly correlated** with X_i, then $R^2_{xw} \neq 0$ and

- What will happen if controlling an irrelevant variable in a regression?*

Relevant Variables and Non-Omitted Variables: Estimate

- Our regression models is still (7.1) and (7.2), but W now is not an irrelevant variable but a **Non-omitted variable**, thus

- Then based on the OVB formula for (7.2),we still have

- The OLS estimator $\hat{\beta}_1$ is still **consistent** whether you include W or not.

Relevant Variables and Non-Omitted Variables: Variance

- Because $\gamma \neq 0$ and $v_i = \gamma W_i + u_i$,thus

- Since we have $Cov(X_i, W_i) = 0$, thus $R^2_{xw} = 0$, then

- It **decrease** the variance of estimator, in other words,it will make the estimate **more precise**.
- **Conclusion**:

Relevant Variables: Omitted Variables

- What about a **Relevant Variables** and correlated with X in our regression model?

- thus the standard definition of **Omitted Variable Bias** if we left it out in our regression model.
- Then based on the OVB formula for (7.2),we still have

- The OLS estimator $\hat{\beta}_1$ is **inconsistent** if you do not include W in the regression model.
- **Conclusion**:

Good Controls v.s Bad Controls

Bad Controls v.s Omitted Variable Bias

- It seems that controlling for more covariates always increases the likelihood that regression estimates have a causal interpretation.
  - often true, but not always.
- eg. Some researchers regressing earnings(Y_i) on schooling(S_i) (and experience) include controls for occupation(O_i). Thus our regression model is

$$Y_i = \beta_0 + \beta_1 S_i + \gamma O_i + u_i$$

  where β_1 is the most of interest coefficient.
- Clearly we can also think of schooling(S_i) affecting the access to higher level occupations(O_i),
  - e.g. you need a Ph.D. to become a university professor. thus

Bad Controls v.s Omitted Variable Bias

- Assume that the true relation is a two equation system: a simultaneous equations system

- In the case, Occupation O_i is an *endogenous variable*.
- As a result, you could not necessarily estimate the first equation by OLS, which means that the estimation of β_1 is not *unbiased* and *consistent*, because of controlling Occupation(O_i).

Bad Controls: Occupation

- Let us come back to the wage premium of college graduation: the conditional expectation.But now we have additional control variable-*occupations*: *white-color* and *blue-olor*
- Two reasonable assumptions:
  1. white-collar jobs, on average, pay more than blue-collar jobs.
  2. graduating college increases the likelihood of a white-collar job.
- **Question 1**: Is occupation an omitted variable in the regression of college degree on wage?
- **Question 2**: Should we control for occupations when considering the effect of college graduation on wages?

Bad Controls: Occupation

- Assume that college degrees are randomly assigned, then we just need to compare the wage difference between workers with college degrees and those without degrees.
- Now we **control** the occupation, which means when we do as follows conditional on occupation:
  - compare degree-earners who chose blue-collar jobs to non-degree-earners who chose blue-collar jobs.
  - or compare degree-earners who chose white-collar jobs to non-degree-earners who chose white-collar jobs.
- Note: the assumption of random degrees says nothing about random job selection.

Bad Controls: Occupation

More formally,

- Y_i denotes i's earnings
- W_i is also a dummy for whether individual i has a white-collar job
- D_i a dummy variable, refers to i's college-graduation status which is randomly assigned, which indicates

- Then

Bad Controls: Occupation

- Because we've assumed D_i is randomly assigned, differences in means yield causal estimates, *i.e.*

Bad Controls: Occupation

- What happens when we estimate the wage-effect of college graduation for white-collar jobs by controlling occupations?

- By introducing a *bad control*, we introduced **selection bias** into a setting that did not have selection bias without controls.

Bad Controls: Occupation

- Specifically,

- The First term: Expected potential non-college earnings, given that potential white collar status associated with college education is equal to 1.
- If the occupational choice between white-collor and blue-collor is randomly assigned, then

$$E\left[\mathbf{Y}_{0i} \mid \mathbf{W}_{1i}=1\right]=E\left[\mathbf{Y}_{0i} \mid \mathbf{W}_{0i}=1\right]$$

- It describes how college graduation changes the composition of the pool of white-collar workers, which in turn change the wage premium between college and high school graduates.
- Even if the true wage causal effect is zero, this selection bias need not be zero.

Bad Controls v.s Omitted Variable Bias

- Bad Controls: Including too many control variables can be counterproductive. They may introduce post-treatment bias when these variables are themselves outcomes of the treatment variable X (essentially functioning as another dependent variable).
- OVB: but if you don't control more variables,you may suffer OVB, which also lead a unbiased and inconsistent estimate.

Bad Controls v.s Omitted Variable Bias

- Bad Controls: Including too many control variables can be counterproductive. They may introduce post-treatment bias when these variables are themselves outcomes of the treatment variable X (essentially functioning as another dependent variable).
- OVB: but if you don't control more variables,you may suffer OVB, which also lead a unbiased and inconsistent estimate.

- "A Hard and Unsolved Problem in Social Sciences" by Gary King(2010), Statistician at Harvard University.
- It is the most "artistic" part of the econometrics in my opinion.

Wrap up

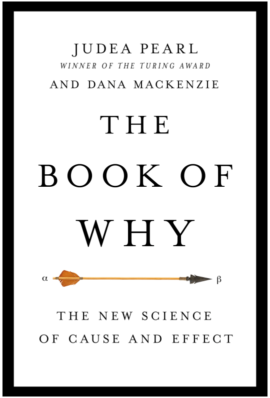
- Which variables should be included on the right-hand side of a regression equation?
  1. **Relevant and Omitted Variables**: These are variables that correlate with both the treatment and the outcome.
     - However, be careful of **bad controls**, as they can introduce more bias.
  2. **Relevant but Non-Omitted Variables**: These are variables that don't correlate with the treatment but do correlate with the outcome.
     - Including these variables may help reduce standard errors.
- Which variables should **NOT be included** on the right-hand side of the equation?
  - Variables that are **irrelevant**.
  - Variables that are **highly correlated with treatment variable** as they may introduce **multicollinearity**.
  - Variables that are **outcome of the treatment variable** as they may introduce **post-treatment bias**.

DAGs and Control Variables

Introduction

- **Directed Acyclic Graphs (DAGs)** graphically illustrates the causal relationships and non-causal associations within a network of random variables.
  - Like **Mind Map**, but more complex for causal inference.
- It can be seen as the other framework to think about the **causality** between variables, besides the **potential outcome framework** we have learned in the second lecture.
  - In my personal opinion, it is a more intuitive and easier way to understand the **causality** between variables.
  - especially, it can help us identify **bad controls** and **omitted variable bias**.

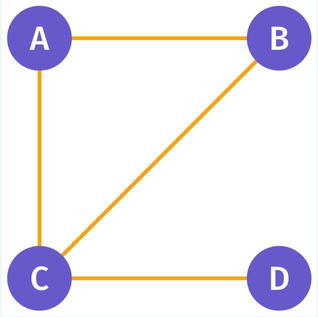
Judea Pearl and Causal DAGs



- Computer Scientist, Turing Award Winner in 2011.
  - *"for fundamental contributions to artificial intelligence through the development of a calculus for probabilistic and causal reasoning".*

Graphs

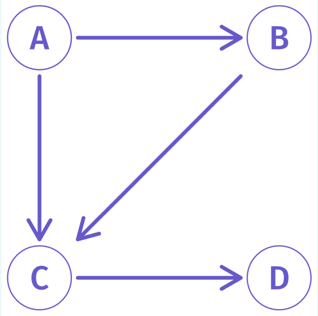
- In graph theory, a graph is a collection of **nodes** connected by **edges**



- Nodes(结点) connected by an edge(边或者连线) are called **adjacent**(邻近).
- **Paths** run along adjacent nodes: $A - B - C$
- The graph above is **undirected**, since the edges don't have direction.

Directed Graphs

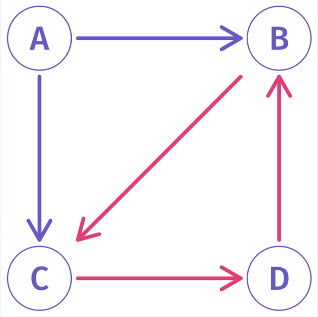
- **Directed** graphs have edges with direction: $A \to B \to C$



- **Ancestors** are nodes that precede a given node in a directed path.
- **Descendants** come after the ancestor node. eg. D is a descendant of C.

Cycles in Graphs

- If a node is its own descendant, then the graph has a **cycle**.

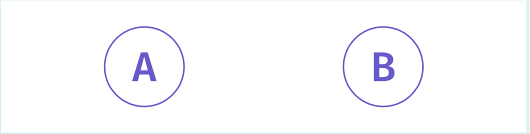


- **Directed Acyclic Graph(DAG)** is the directed graph which does not have any cycles.

DAG: Building blocks

- Focusing the relationship between variables(nodes)
    - **Dependent** or **Independent**

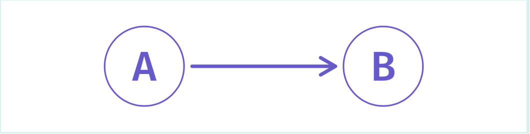
1. **Two unconnected nodes**



- A and B are independent—no link between the nodes.

DAG: Building blocks

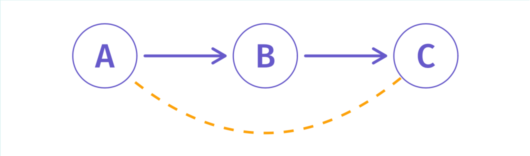
2. **Two connected nodes**



- There is clear (causal) dependence: *A is a cause of B*.

DAG: Building blocks

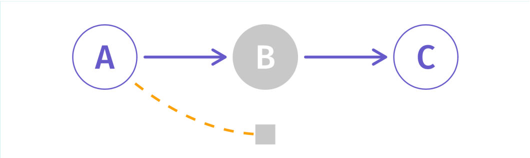
3. **Chains**



- A and B are dependent.
- B and C are dependent.
- Then are A and C dependent?
    - Mostly yes.through the chain A-B-C.

DAG: Building blocks

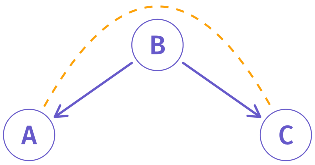
3. **Chains with conditions**



- **Quesiton**: how does conditioning on B affect the association between A and C?
- **Answer**: It breaks the chain between A and C, thus A and C are independent.

DAG: Building blocks

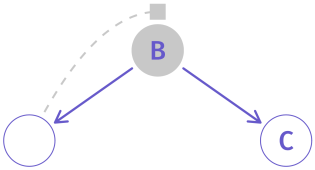
4. **Forks or Confounds**



- A and C are usually associated in forks.
  - B induces changes in A
  - B also induces changes in C
- A and C are **associated** due to their common cause, typically OVB.

DAG: Building blocks

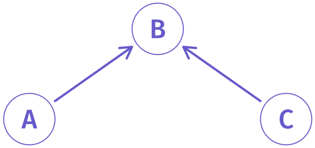
4. **Blocked forks**



- **Question**: What happens when we condition on B?
- **Answer**: Conditioning on B makes A and C independent.
  - A and C are only associated due to their common cause B.
  - When we shutdown (hold constant) this common cause (B), there is no way for A and C to associate.

DAG: Building blocks

5. **Immoralities or Colliders**



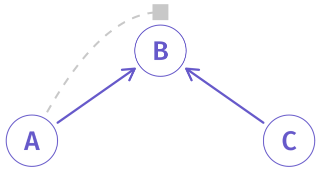
- An **immorality** occurs when two nodes share a child without being otherwise connected

$$A \rightarrow B \leftarrow C$$

- The child (here:B) at the center of this immorality is called a **collider**.

DAG: Building blocks

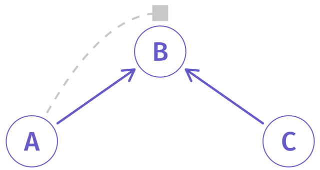
5. **Immoralities or Colliders**



- **Question**: Are A and C independent?
- **Answer**: Yes. A $\perp$ C.
  - Causal effects flow from A and C and stop there.
  - Neither A nor C is a descendant of the other.
  - A and C do not share any common causes.

DAG: Building blocks

5. **Immoralities with conditions**

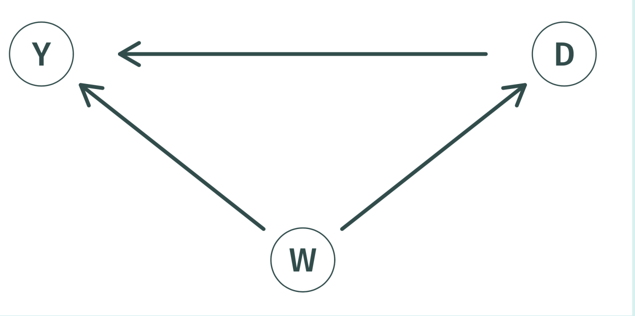


- **Question**: What happens when we condition on B?
- **Answer**: We **unblock** or **open** the previously blocked (closed) path.
- While A and C are independent, they are conditionally dependent.
- When you condition on a collider, you open up the path.

DAGs: Blocked Paths

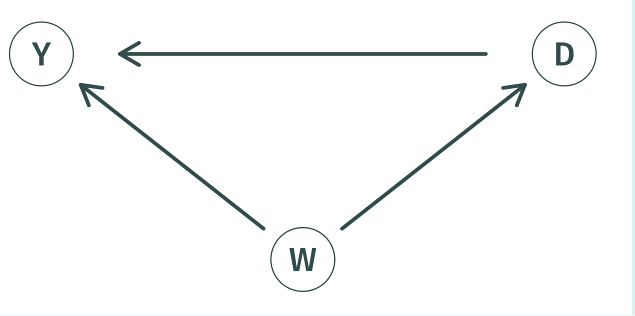
- A path between X and Y is **blocked** by conditioning on a set of variables Z if either of the following statements is true:
  1. On the path, there is a **chain** $(\cdots \to W \to \dots)$ or a **fork** $(\cdots \leftarrow W \to \dots)$, and we condition on W $(W \in Z)$.
  2. On the path, there is a **collider** $(\cdots \to W \leftarrow \dots)$, and we do **NOT** condition on W $(W \notin Z)$ or any of its descendants $(de(W) \not\subseteq Z)$.

OVB in a DAG



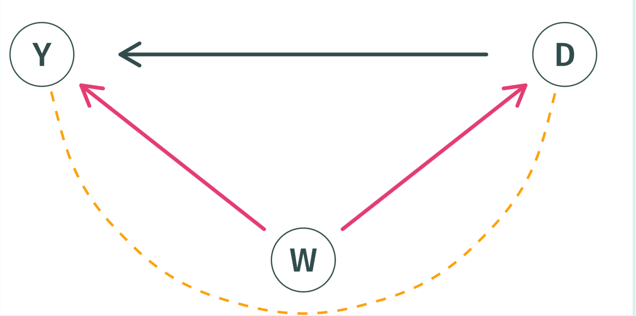
- **Nodes** are random variables.
  - Y is the dependent variable
  - D is the treatment variable
  - W is the Omitted variable

OVB in a DAG



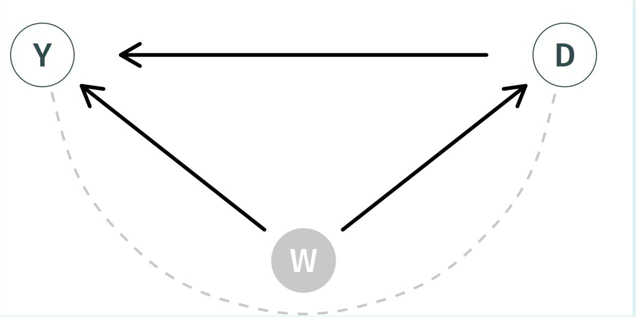
- **Edges** depict causal links.
- **Causality** flows in the direction of the arrows.
- Both connections and non-connections matter!

OVB in a DAG



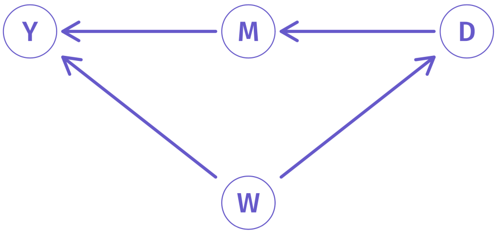
- There are two pathways from D to Y:
  1. The path from D to Y is our casual relationship of interest.
  2. The path $Y \leftarrow W \rightarrow D$ creates non-causal association between D and Y.

OVB in a DAG



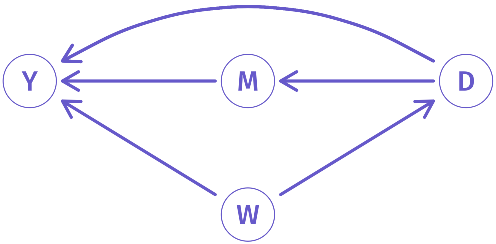
- To shut down this pathway creating a non-causal association, we have to **control/balance/adjust** on W.

Mediation in a DAG



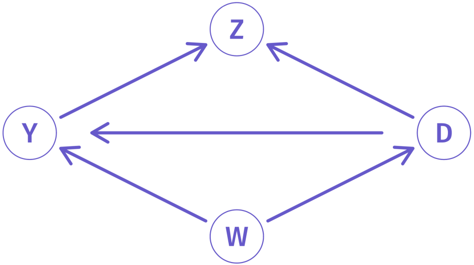
- **Question**: How to control variables to isolate the causal effect of D on Y?
- **Answer**: Control on W and Don't control on M

Partial Mediation in a DAG



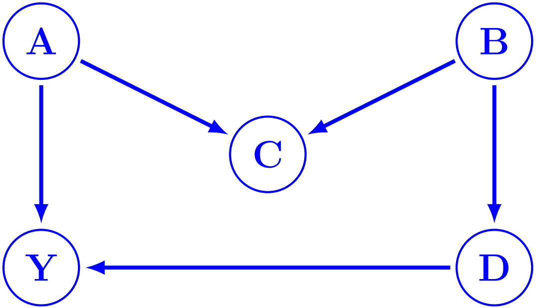
- **Question**: How to control variables to isolate the causal effect of D on Y?
- **Answer**: Control on W and Don't control on M, either.
  - Why? Because our causal effect of interest is an aggregate effect, not a partial effect.
  - Or you could underestimate the effect of D on Y.

DAGs: Example 4 Non-mediator descendants



- **Question**: How to control variables to isolate the causal effect of D on Y?
- **Answer**: Control on W and Don't control on Z.
    - here Z is a collider variable, controlling on it will open the backdoor path.

DAGs: Example 5 M-Bias



- **Question**: How to control variables to isolate the causal effect of D on Y?
- **Answer**: Control on W and Don't control on C.
  - here C is a collider variable, controlling on it will open the backdoor path.
- **Question**: How about controlling on A and B?
- **Answer**:

DAG Applications: Going to College on Earnings

- Our Simple OLS Regression:

$$\log(\mathbf{Y}) = \beta_0 + \beta_1 \mathbf{D} + \varepsilon$$

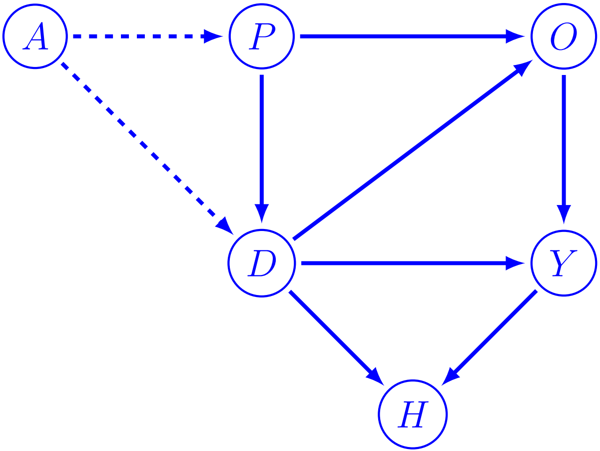
  - where Y is earnings, D is a dummy variable for going to college, ε is the error term. The coefficient β_1 represents the causal effect of going to college on earnings or *the return to college*.
- However, the OLS estimator is biased and inconsistent because of the OVB and confounding factors.
- There are some other factors that affect the wage and schooling as well.
  - Some are observed: **Parental education(P)**, **Occupation(O)**, **Social Network(N)**, etc.
  - Some are unobserved: **Ability from genetics(A)**,etc.

DAG Applications: Going to College on Earnings

- Suppose
  - Children's Schooling(D) can affect Occupation(O), Earnings(Y) and Health(H).
  - Parents' SES(P) can affect Children's Schooling(D) and Occupation(O).
  - Children's Earnings(Y) can affect their Health(H).
  - Children's Occupation(O) can affect their Earnings(Y).
  - Ability from genetics(A) of family can affect Children's schooling(D) and Parents' SES(P).

DAG Applications: Going to College on Earnings

- Then the DAG is:



DAG Applications: Going to College on Earnings

- The path from D to Y is our causal relationship of interest. Then list all the paths from D to Y.
  - $D \to Y$(causal relationship of interest)
  - $D \leftarrow P \to O \to Y$(Partial Mediator)
  - $D \leftarrow P \to O \to Y$(confounder)
  - $D \leftarrow A \to P \to O \to Y$(confounder)
  - $D \to H \leftarrow Y$(collider)
- **Question**: Which variable should we control on?

Limitations of DAGs

- DAGs are typically depicted without “noise variables” or disturbances.
- However, these disturbances still exist—they’re just “outside of the model.”
- Simultaneity defines causality as unidirectional and disallows cycles.
- Dynamics allow a variable to somewhat influence itself, especially in time series.
- Essentially, this is a form of logical deduction that proves to be very useful.

Final Remarks about DAGs

- DAGs are a powerful tool for causal inference and can help us identify **bad controls** and **omitted variable bias**.
- However, they heavily depends on the **knowledge of the causal relationships** and **assumptions** we made.
  - Knowledge of the causal relationships: from theory, model, observation, experience, priors, etc.
  - The best part of DAGs is that it makes our assumptions and mechanisms in our regression model **explicit**.
- DAGs can not be a replacement for statistical models, but rather a complement to them.

Some practical tips about Control Variables

Practical Tip 1

- **Use theory and prior research to guide control selection:**

Practical Tip 2

- **Sensitivity analysis**:

Practical Tip 3

- **Remember the fundamental trade-off:**