Lecture 7: Assessing Regression Studies(I)

Introduction to Econometrics,Spring 2026

Zhaopeng Qu

Business School,Nanjing University

April 29 2026

1. Review of Previous Lectures
2. Accessing Regression Studies: Introduction
3. Omitted Variable Bias(OVB) and Control Variables
4. DAGs and Control Variables
5. Some practical tips about Control Variables

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: the statistical inferences about causal effects are valid for the population and setting **being studied**.
  - External validity: the statistical inferences can be generalized from the population and setting studied to **other populations and settings**.
- 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 10 times than the external validity in most studies.
- Of course, it is not to say that we should ignore the external validity, which determines the generalizability of the results to other populations and settings.
- But we should focus on the internal validity **first**. 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

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

- Internal validity has **three components**:

1. The estimators of β_1 should be unbiased and consistent. This is the most critical aspect.
2. Hypothesis tests and confidence intervals should have the desired significance level (at least 5% significant).
3. The value of β_1 should be large enough to be meaningful or economically significant.

Threats to Internal Validity

- Threats to internal validity:
  1. **Omitted Variables**
  2. **Misspecification**
  3. **Measurement Error**
  4. **Simultaneous Causality**
  5. **Missing Data and Sample Selection**
  6. **Heteroskedasticity and/or Correlated error terms**
  7. **Significant coefficients or marginal effects**
- 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:
  - True model:

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

    where $E(u_i|X_i, W_i) = 0$
  - Observed model:

$$Y_i = \beta_0 + \beta_1 X_i + v_i$$

    where $v_i = \gamma W_i + u_i$

OVB Review

- Then we have the OLS estimator of β_1 as follows

$$plim\hat{\beta}_1 = \beta_1 + \gamma \frac{Cov(X_i, W_i)}{VarX_i}$$

- An omitted variable W_i leads to an inconsistent OLS estimate of the causal effect of X_i if **both**
  - W_i is related to X, thus $Cov(X_i, W_i) \neq 0$
  - W_i has some effect on Y_i, thus $\gamma \neq 0$
- 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 covariates variables, or 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

- **Multiple OLS model:**

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

- **The OLS estimator of β_j based on FWL theorem:**

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

- **The asymptotic OLS estimator of β_j**

$$\text{plim}\, \hat{\beta}_j = \frac{\text{Cov}(\tilde{X}_{ij}, Y_i)}{\text{Var}(\tilde{X}_{ij})}$$

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

$$X_{ij} = \hat{\gamma}_0 + \hat{\gamma}_1 X_{i1} + \hat{\gamma}_2 X_{i2} + ... + \hat{\gamma}_{j-1} X_{i,j-1} + \hat{\gamma}_{j+1} X_{i,j+1} + ... + \hat{\gamma}_k X_{ik} + \tilde{X}_{ij}$$

Recall: Unbiasedness of OLS Estimators

- The auxiliary-regression residual $\tilde{X}_{ij}$ satisfies three orthogonality identities:
  - (a) $\sum_{i=1}^{n} \tilde{X}_{ij} = 0$;
  - (b) $\sum_{i=1}^{n} \tilde{X}_{ij} X_{i\ell} = 0$ for every $\ell \neq j$;
  - (c) $\sum_{i=1}^{n} \tilde{X}_{ij} X_{ij} = \sum_{i=1}^{n} \tilde{X}_{ij}^2$. These three identities are why every β_ℓ with $\ell \neq j$ drops out below — this is **Frisch–Waugh–Lovell (FWL)** at work.
- Based on the multiple OLS estimators in ,we have

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

Recall: Unbiasedness of OLS Estimators

- The auxiliary-regression residual $\tilde{X}_{ij}$ satisfies three orthogonality identities:
  - (a) $\sum_{i=1}^{n} \tilde{X}_{ij} = 0$;
  - (b) $\sum_{i=1}^{n} \tilde{X}_{ij} X_{i\ell} = 0$ for every $\ell \neq j$;
  - (c) $\sum_{i=1}^{n} \tilde{X}_{ij} X_{ij} = \sum_{i=1}^{n} \tilde{X}_{ij}^2$. These three identities are why every β_ℓ with $\ell \neq j$ drops out below — this is **Frisch–Waugh–Lovell (FWL)** at work.
- Based on the multiple OLS estimators in ,we have

$$\hat{\boldsymbol{\beta}}_j = \frac{\sum_{i=1}^{n} \tilde{X}_{ij} Y_i}{\sum_{i=1}^{n} \tilde{X}_{ij}^2} = \frac{\sum_{i=1}^{n} \tilde{X}_{ij} (\boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 X_{i1} + \boldsymbol{\beta}_2 X_{i2} + \ldots + \boldsymbol{\beta}_k X_{ik} + u_i)}{\sum_{i=1}^{n} \tilde{X}_{ij}^2}$$

Recall: Unbiasedness of OLS Estimators

- The auxiliary-regression residual $\tilde{X}_{ij}$ satisfies three orthogonality identities:
  (a) $\sum_{i=1}^{n} \tilde{X}_{ij} = 0$;
  (b) $\sum_{i=1}^{n} \tilde{X}_{ij} X_{i\ell} = 0$ for every $\ell \neq j$;
  (c) $\sum_{i=1}^{n} \tilde{X}_{ij} X_{ij} = \sum_{i=1}^{n} \tilde{X}_{ij}^2$. These three identities are why every β_ℓ with $\ell \neq j$ drops out below — this is **Frisch–Waugh–Lovell (FWL)** at work.
- Based on the multiple OLS estimators in ,we have

$$\hat{\boldsymbol{\beta}}_j = \frac{\sum_{i=1}^{n} \tilde{X}_{ij} Y_i}{\sum_{i=1}^{n} \tilde{X}_{ij}^2} = \frac{\sum_{i=1}^{n} \tilde{X}_{ij} (\boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 X_{i1} + \boldsymbol{\beta}_2 X_{i2} + \ldots + \boldsymbol{\beta}_k X_{ik} + u_i)}{\sum_{i=1}^{n} \tilde{X}_{ij}^2}$$

$$= \frac{\boldsymbol{\beta}_j \sum_{i=1}^{n} \tilde{X}_{ij} X_{ij} + \sum_{i=1}^{n} \tilde{X}_{ij} u_i}{\sum_{i=1}^{n} \tilde{X}_{ij}^2}$$

Recall: Unbiasedness of OLS Estimators

- The auxiliary-regression residual $\tilde{X}_{ij}$ satisfies three orthogonality identities:
  (a) $\sum_{i=1}^{n} \tilde{X}_{ij} = 0$;
  (b) $\sum_{i=1}^{n} \tilde{X}_{ij} X_{i\ell} = 0$ for every $\ell \neq j$;
  (c) $\sum_{i=1}^{n} \tilde{X}_{ij} X_{ij} = \sum_{i=1}^{n} \tilde{X}_{ij}^2$. These three identities are why every β_ℓ with $\ell \neq j$ drops out below — this is **Frisch–Waugh–Lovell (FWL)** at work.
- Based on the multiple OLS estimators in ,we have

$$\hat{\boldsymbol{\beta}}_j = \frac{\sum_{i=1}^{n} \tilde{X}_{ij} Y_i}{\sum_{i=1}^{n} \tilde{X}_{ij}^2} = \frac{\sum_{i=1}^{n} \tilde{X}_{ij} (\boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 X_{i1} + \boldsymbol{\beta}_2 X_{i2} + \ldots + \boldsymbol{\beta}_k X_{ik} + u_i)}{\sum_{i=1}^{n} \tilde{X}_{ij}^2}$$

$$= \frac{\boldsymbol{\beta}_j \sum_{i=1}^{n} \tilde{X}_{ij} X_{ij} + \sum_{i=1}^{n} \tilde{X}_{ij} u_i}{\sum_{i=1}^{n} \tilde{X}_{ij}^2}$$

$$= \boldsymbol{\beta}_j + \frac{\sum_{i=1}^{n} \tilde{X}_{ij} u_i}{\sum_{i=1}^{n} \tilde{X}_{ij}^2}$$

Recall: Unbiasedness of OLS Estimators

- The auxiliary-regression residual $\tilde{X}_{ij}$ satisfies three orthogonality identities:
  (a) $\sum_{i=1}^{n} \tilde{X}_{ij} = 0$;
  (b) $\sum_{i=1}^{n} \tilde{X}_{ij} X_{i\ell} = 0$ for every $\ell \neq j$;
  (c) $\sum_{i=1}^{n} \tilde{X}_{ij} X_{ij} = \sum_{i=1}^{n} \tilde{X}_{ij}^2$. These three identities are why every β_ℓ with $\ell \neq j$ drops out below — this is **Frisch–Waugh–Lovell (FWL)** at work.
- Based on the multiple OLS estimators in ,we have

$$\hat{\boldsymbol{\beta}}_j = \frac{\sum_{i=1}^{n} \tilde{X}_{ij} Y_i}{\sum_{i=1}^{n} \tilde{X}_{ij}^2} = \frac{\sum_{i=1}^{n} \tilde{X}_{ij} (\boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 X_{i1} + \boldsymbol{\beta}_2 X_{i2} + \ldots + \boldsymbol{\beta}_k X_{ik} + u_i)}{\sum_{i=1}^{n} \tilde{X}_{ij}^2}$$

$$= \frac{\boldsymbol{\beta}_j \sum_{i=1}^{n} \tilde{X}_{ij} X_{ij} + \sum_{i=1}^{n} \tilde{X}_{ij} u_i}{\sum_{i=1}^{n} \tilde{X}_{ij}^2}$$

$$= \boldsymbol{\beta}_j + \frac{\sum_{i=1}^{n} \tilde{X}_{ij} u_i}{\sum_{i=1}^{n} \tilde{X}_{ij}^2}$$

Variance of OLS Estimators

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

$$\mathrm{Var}(u_i|X_{1i}, X_{2i}..., X_{ki}) = \mathrm{Var}(u_i) = \sigma_u^2$$

- **Note**: All variance operations below are understood as **conditional variances** $\mathrm{Var}(\cdot \mid X_1, \ldots, X_n)$. Under i.i.d. sampling $\{(Y_i, X_i)\}$, conditioning on X lets us treat $\tilde{X}_{ij}$ as **known constants** instead of random variables.
- Actually, the unconditional variance follows by the **Law of Total Variance**, and the two will converge to the same value *in large samples*.

Variance of OLS Estimators

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

$$\text{Var}(\hat{\beta}_j)$$

Variance of OLS Estimators

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

$$\text{Var}(\hat{\beta}_j) = \text{Var}\left(\boldsymbol{\beta}_j + \frac{\sum_{i=1}^{n} \tilde{X}_{ij} u_i}{\sum_{i=1}^{n} \tilde{X}_{ij}^2}\right)$$

Variance of OLS Estimators

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

$$\begin{aligned}\mathrm{Var}(\hat{\beta}_j) &= \mathrm{Var}\left(\boldsymbol{\beta}_j + \frac{\sum_{i=1}^{n} \tilde{X}_{ij} u_i}{\sum_{i=1}^{n} \tilde{X}_{ij}^2}\right) \\ &= \mathrm{Var}(\boldsymbol{\beta}_j) + \mathrm{Var}\left(\frac{\sum_{i=1}^{n} \tilde{X}_{ij} u_i}{\sum_{i=1}^{n} \tilde{X}_{ij}^2}\right)\end{aligned}$$

Variance of OLS Estimators

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

$$
\begin{aligned}
\mathrm{Var}(\hat{\beta}_j) &= \mathrm{Var}\left(\boldsymbol{\beta}_j + \frac{\sum_{i=1}^{n} \tilde{X}_{ij} u_i}{\sum_{i=1}^{n} \tilde{X}_{ij}^2}\right) \\
&= \mathrm{Var}(\boldsymbol{\beta}_j) + \mathrm{Var}\left(\frac{\sum_{i=1}^{n} \tilde{X}_{ij} u_i}{\sum_{i=1}^{n} \tilde{X}_{ij}^2}\right) \\
&= \frac{\sum_{i=1}^{n} \tilde{X}_{ij}^2 \, \mathrm{Var}(u_i)}{\left(\sum_{i=1}^{n} \tilde{X}_{ij}^2\right)^2} \quad \text{for i.i.d sample and constant } \tilde{X}_{ij}
\end{aligned}
$$

Variance of OLS Estimators

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

$$
\begin{aligned}
\operatorname{Var}(\hat{\beta}_j) &= \operatorname{Var}\left(\boldsymbol{\beta}_j + \frac{\sum_{i=1}^{n} \tilde{X}_{ij} u_i}{\sum_{i=1}^{n} \tilde{X}_{ij}^2}\right) \\
&= \operatorname{Var}(\boldsymbol{\beta}_j) + \operatorname{Var}\left(\frac{\sum_{i=1}^{n} \tilde{X}_{ij} u_i}{\sum_{i=1}^{n} \tilde{X}_{ij}^2}\right) \\
&= \frac{\sum_{i=1}^{n} \tilde{X}_{ij}^2 \operatorname{Var}(u_i)}{\left(\sum_{i=1}^{n} \tilde{X}_{ij}^2\right)^2} \text{ for i.i.d sample and constant } \tilde{X}_{ij} \\
&= \frac{\sum_{i=1}^{n} \tilde{X}_{ij}^2 \sigma_u^2}{\left(\sum_{i=1}^{n} \tilde{X}_{ij}^2\right)^2} \text{ for homoskedasticity}
\end{aligned}
$$

Variance of OLS Estimators

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

$$
\begin{aligned}
\mathrm{Var}(\hat{\beta}_j) &= \mathrm{Var}\left(\boldsymbol{\beta}_j + \frac{\sum_{i=1}^{n} \tilde{X}_{ij} u_i}{\sum_{i=1}^{n} \tilde{X}_{ij}^2}\right) \\
&= \mathrm{Var}(\boldsymbol{\beta}_j) + \mathrm{Var}\left(\frac{\sum_{i=1}^{n} \tilde{X}_{ij} u_i}{\sum_{i=1}^{n} \tilde{X}_{ij}^2}\right) \\
&= \frac{\sum_{i=1}^{n} \tilde{X}_{ij}^2 \,\mathrm{Var}(u_i)}{\left(\sum_{i=1}^{n} \tilde{X}_{ij}^2\right)^2} \text{ for i.i.d sample and constant } \tilde{X}_{ij} \\
&= \frac{\sum_{i=1}^{n} \tilde{X}_{ij}^2 \sigma_u^2}{\left(\sum_{i=1}^{n} \tilde{X}_{ij}^2\right)^2} \text{ for homoskedasticity} \\
&= \frac{\sigma_u^2}{\sum_{i=1}^{n} \tilde{X}_{ij}^2}
\end{aligned}
$$

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

- The R-Squared of this partitioned regression for the variable of interest X_j on other regressors is

$$R_j^2 = 1 - \frac{SSR_j}{TSS_j}$$

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

- The R-Squared of this partitioned regression for the variable of interest X_j on other regressors is

$$R_j^2 = 1 - \frac{SSR_j}{TSS_j}$$

$$\Rightarrow SSR_j = (1 - R_j^2) \times TSS_j$$

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

- The R-Squared of this partitioned regression for the variable of interest X_j on other regressors is

$$R_j^2 = 1 - \frac{SSR_j}{TSS_j}$$

$$\Rightarrow SSR_j = (1 - R_j^2) \times TSS_j$$

$$\Rightarrow \sum_{i=1}^{n} \tilde{X}_{ij}^2 = (1 - R_j^2) \sum_{i=1}^{n} (X_{ji} - \bar{X}_j)^2 = (1 - R_j^2)(n-1)s_j^2$$

- 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.
- Then the variance of $\hat{\beta}_j$ under homoskedasticity is

$$\text{Var}\left(\hat{\beta}_j\right) = \frac{\sigma_u^2}{\sum_{i=1}^{n} \tilde{X}_{ij}^2}$$

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

- The R-Squared of this partitioned regression for the variable of interest X_j on other regressors is

$$R_j^2 = 1 - \frac{SSR_j}{TSS_j}$$

$$\Rightarrow SSR_j = (1 - R_j^2) \times TSS_j$$

$$\Rightarrow \sum_{i=1}^{n} \tilde{X}_{ij}^2 = (1 - R_j^2) \sum_{i=1}^{n} (X_{ji} - \bar{X}_j)^2 = (1 - R_j^2)(n-1)s_j^2$$

- 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.
- Then the variance of $\hat{\beta}_j$ under homoskedasticity is

$$\text{Var}\left(\hat{\beta}_j\right) = \frac{\sigma_u^2}{\sum_{i=1}^{n} \tilde{X}_{ij}^2} = \frac{\sigma_u^2}{(n-1)s_j^2(1 - R_j^2)}$$

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

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

$$\text{Var}\left(\hat{\beta}_j\right) = \sigma^2_{\hat{\beta}_j} = \frac{\sigma^2_u}{(n-1)s^2_j(1-R^2_j)}$$

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

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

$$\text{Var}\left(\hat{\beta}_j\right) = \sigma^2_{\hat{\beta}_j} = \frac{\sigma_u^2}{(n-1)s_j^2(1-R_j^2)}$$

- **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

$$\text{Var}\left(\hat{\beta}_j\right) = \sigma^2_{\hat{\beta}_j} = \frac{\sigma_u^2}{(n-1)s_j^2(1-R_j^2)}$$

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

Factors	symbols	$\text{Var}\left(\hat{\beta}_j\right)$
the variance of u_i	$\sigma_u^2 \uparrow$	$\uparrow$

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

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

$$\text{Var}\left(\hat{\beta}_j\right) = \sigma^2_{\hat{\beta}_j} = \frac{\sigma_u^2}{(n-1)s_j^2(1-R_j^2)}$$

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

Factors	symbols	$\text{Var}\left(\hat{\beta}_j\right)$
the variance of u_i	$\sigma_u^2 \uparrow$	$\uparrow$
the sample variance of X_j	$s_j^2 \uparrow$	$\downarrow$

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

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

$$\text{Var}\left(\hat{\beta}_j\right) = \sigma^2_{\hat{\beta}_j} = \frac{\sigma_u^2}{(n-1)s_j^2(1-R_j^2)}$$

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

Factors	symbols	$\text{Var}\left(\hat{\beta}_j\right)$
the variance of u_i	$\sigma_u^2 \uparrow$	$\uparrow$
the sample variance of X_j	$s_j^2 \uparrow$	$\downarrow$
the R_j^2	$R_j^2 \uparrow$	$\uparrow$

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

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

$$\text{Var}\left(\hat{\beta}_j\right) = \sigma^2_{\hat{\beta}_j} = \frac{\sigma_u^2}{(n-1)s_j^2(1-R_j^2)}$$

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

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

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: W has a **NONZERO** partial effect on the dependent variable Y.
  - Irrelevant Variables: W has a **ZERO** partial effect on the dependent variable Y.

  2. Whether W is correlated with X or not:
  - Uncorrelated Control Variables: W is **NOT** correlated with X.
  - Correlated Control Variables: W is correlated with X.
  - Highly-Correlated Control Variables: W is **highly** correlated with X.

Control Variables: W

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

W	Irrelevant to Y	Relevant to Y
Uncorrelated with X		
Correlated with X		
Highly-Correlated with X		

Control Variables: W

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

W	Irrelevant to Y	Relevant to Y
Uncorrelated with X **Correlated** with X **Highly-Correlated** with X	**Irrelevant Variables**	

Control Variables: W

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

W	Irrelevant to Y	Relevant to Y
Uncorrelated with X **Correlated** with X **Highly-Correlated** with X	**Irrelevant Variables**	**Non-Omitted Variables**

Control Variables: W

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

W	Irrelevant to Y	Relevant to Y
Uncorrelated with X	**Irrelevant Variables**	**Non-Omitted Variables**
Correlated with X	**Irrelevant Variables**	
Highly-Correlated with X		

Control Variables: W

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

W	Irrelevant to Y	Relevant to Y
Uncorrelated with X	**Irrelevant Variables**	**Non-Omitted Variables**
Correlated with X	**Irrelevant Variables**	**Omitted Variables**
Highly-Correlated with X		

Control Variables: W

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

W	Irrelevant to Y	Relevant to Y
Uncorrelated with X	**Irrelevant Variables**	**Non-Omitted Variables**
Correlated with X	**Irrelevant Variables**	**Omitted Variables**
Highly-Correlated with X	**(Worse) Irrelevant Variables**	

Control Variables: W

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

W	Irrelevant to Y	Relevant to Y
Uncorrelated with X	**Irrelevant Variables**	**Non-Omitted Variables**
Correlated with X	**Irrelevant Variables**	**Omitted Variables**
Highly-Correlated with X	**(Worse) Irrelevant Variables**	**Multicollinearity and OVB**

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

$$\gamma = 0$$

- Then the model excluding W is

$$Y_i = \tilde{\beta}_0 + \tilde{\beta}_1 X_i + v_i \tag{7.2}$$

  - 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

$$plim\hat{\tilde{\beta}}_1 = \tilde{\beta}_1 + \gamma \frac{Cov(X_i, W_i)}{VarX_i} = \tilde{\beta}_1 = \beta_1$$

- 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**

$$Var(\hat{\beta}_1) = \frac{\sigma_v^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2(1 - R_{xw}^2)} \tag{7.3}$$

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

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

$$Var(\hat{\tilde{\beta}}) = \frac{\sigma_v^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \tag{7.4}$$

Irrelevant Variables: Variance

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

$$Var(v_i) = Var(u_i) \Rightarrow \sigma_v^2 = \sigma_u^2$$

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

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

- **Then**

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

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

$$Var(\hat{\beta}_j) = \frac{\sigma_u^2}{(n-1)s_j^2(1-R_j^2)}$$

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

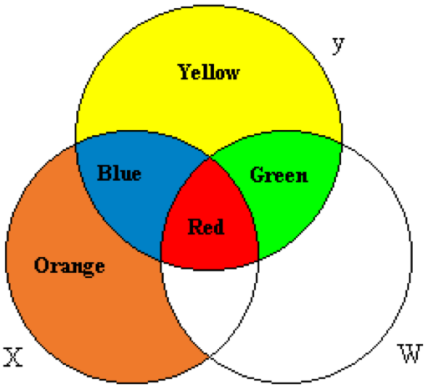
Multicollinearity

- 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:

$$VIF_j = \frac{1}{1 - R_j^2}$$

  - 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

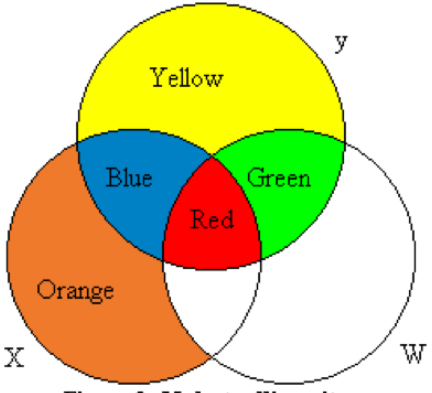


Figure 3a Modest collinearity

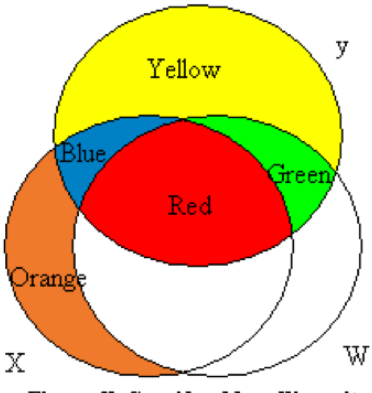


Figure 3b Considerable collinearity

Venn Diagrams for Multicollinearity

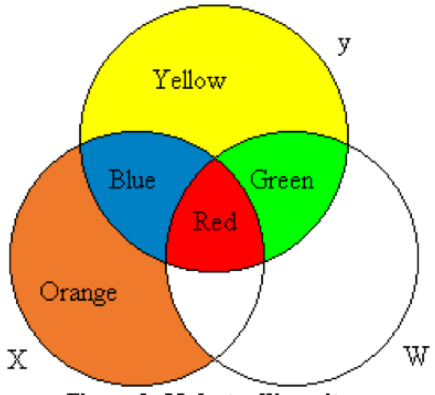


Figure 3a Modest collinearity

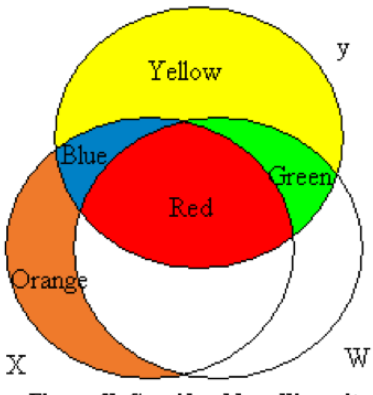


Figure 3b Considerable collinearity

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

Irrelevant Variables: Wrap Up

- Then we have
  1. if irrelevant variable W_i is **not** correlated with X_i, then $R^2_{xw} = 0$ and
  $$Var(\hat{\beta}_1) = Var(\hat{\tilde{\beta}}_1)$$
  2. if irrelevant variable W_i is **correlated** with X_i, then $R^2_{xw} \neq 0$ and
  $$Var(\hat{\beta}_1) > Var(\hat{\tilde{\beta}}_1)$$
  3. if irrelevant variable W_i is **highly correlated** with X_i, then $R^2_{xw} \neq 0$ and
  $$Var(\hat{\beta}_1) >> Var(\hat{\tilde{\beta}}_1)$$
- What will happen if controlling an **irrelevant** variable in a regression?*
  1. The OLS estimator is still unbiased and consistent.
  2. It increase the variance of estimator, in other words, make the estimate **less precise**.
  3. If multicollinearity exists, it will make the estimate **very imprecise**.
- **Conclusion**: *we should avoid to put irrelevant variables into our 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

$$Cov(X_i, W_i) = 0 \ \& \ \gamma \neq 0$$

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

$$plim\hat{\tilde{\beta}}_1 = \tilde{\beta}_1 + \gamma \frac{Cov(X_i, W_i)}{VarX_i} = \tilde{\beta}_1$$

- 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**

$$Var(u_i) \leq Var(v_i) \Rightarrow \sigma_u^2 \leq \sigma_v^2$$

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

$$Var(\hat{\beta}_1) \leq Var(\hat{\tilde{\beta}}_1)$$

- **It decrease the variance of estimator, in other words,it will make the estimate more precise.**
- **Conclusion: we should always put Relevant but Non-Omitted Variables into our regression.**

Relevant Variables: Omitted Variables

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

$$Cov(X_i, W_i) \neq 0 \ \& \ \gamma \neq 0$$

- 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

$$plim\hat{\tilde{\beta}}_1 = \tilde{\beta}_1 + \gamma \frac{Cov(X_i, W_i)}{VarX_i} \neq \tilde{\beta}_1$$

- The OLS estimator $\hat{\beta}_1$ is inconsistent if you do not include W in the regression model.
- **Conclusion**: It seems that we should always put Omitted Variables into our regression.
- In reality, the answer is that it depends on what kind of omitted variables we are dealing with: Good or Bad.

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

$$O_i = \lambda_0 + \lambda_1 S_i + e_i$$

Bad Controls v.s Omitted Variable Bias

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

$$Y_i = \beta_0 + \beta_1 S_i + \gamma O_i + e_i$$
$$O_i = \lambda_0 + \lambda_1 S_i + u_i$$

- In the case, Occupation O_i is an endogenous variable.
- As a result, OLS applied to the first equation does not generally recover a causal effect of schooling once occupation is included: the estimator of β_1 can be biased and inconsistent, because Occupation (O_i) is endogenous.

Bad Controls: Occupation

- Let us come back to the wage premium of college graduation. But now we have additional control variable-occupations: *white-collar* and *blue-collar*
- 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 randomly assigned degrees says nothing about randomly assigned jobs.

Bad Controls: Occupation

More formally,

- $\mathbf{Y}_i$ denotes i's earnings
- $\mathbf{W}_i$ is also a dummy for whether individual i has a white-collar job
- $\mathbf{D}_i$ a dummy variable, refers to i's college-graduation status which is randomly assigned, which indicates

$$\{Y_1, Y_0 \perp D\} \; and \; \{W_1, W_0 \perp D\}$$

- Then

$$\mathbf{Y}_i = \mathbf{D}_i \mathbf{Y}_{1i} + (1 - \mathbf{D}_i) \mathbf{Y}_{0i}$$
$$\mathbf{W}_i = \mathbf{D}_i \mathbf{W}_{1i} + (1 - \mathbf{D}_i) \mathbf{W}_{0i}$$

Bad Controls: Occupation

- Because we've assumed $\mathbf{D}_i$ is randomly assigned, differences in means yield causal estimates, *i.e.*

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

Bad Controls: Occupation

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

Bad Controls: Occupation

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

$$E\left[\mathbf{Y}_i \mid \mathbf{W}_i = 1,\, \mathbf{D}_i = 1\right] - E\left[\mathbf{Y}_i \mid \mathbf{W}_i = 1,\, \mathbf{D}_i = 0\right]$$

Bad Controls: Occupation

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

$$
\begin{aligned}
&E\left[\mathbf{Y}_i \mid \mathbf{W}_i = 1,\, \mathbf{D}_i = 1\right] - E\left[\mathbf{Y}_i \mid \mathbf{W}_i = 1,\, \mathbf{D}_i = 0\right] \\
&= E\left[\mathbf{Y}_{1i} \mid \mathbf{W}_{1i} = 1,\, \mathbf{D}_i = 1\right] - E\left[\mathbf{Y}_{0i} \mid \mathbf{W}_{0i} = 1,\, \mathbf{D}_i = 0\right]
\end{aligned}
$$

Bad Controls: Occupation

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

$$
\begin{aligned}
& E\left[\mathbf{Y}_i \mid \mathbf{W}_i = 1,\, \mathbf{D}_i = 1\right] - E\left[\mathbf{Y}_i \mid \mathbf{W}_i = 1,\, \mathbf{D}_i = 0\right] \\
&= E\left[\mathbf{Y}_{1i} \mid \mathbf{W}_{1i} = 1,\, \mathbf{D}_i = 1\right] - E\left[\mathbf{Y}_{0i} \mid \mathbf{W}_{0i} = 1,\, \mathbf{D}_i = 0\right] \\
&= E\left[\mathbf{Y}_{1i} \mid \mathbf{W}_{1i} = 1\right] - E\left[\mathbf{Y}_{0i} \mid \mathbf{W}_{0i} = 1\right]
\end{aligned}
$$

Bad Controls: Occupation

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

$$
\begin{aligned}
& E\left[\mathbf{Y}_i \mid \mathbf{W}_i = 1,\, \mathbf{D}_i = 1\right] - E\left[\mathbf{Y}_i \mid \mathbf{W}_i = 1,\, \mathbf{D}_i = 0\right] \\
&= E\left[\mathbf{Y}_{1i} \mid \mathbf{W}_{1i} = 1,\, \mathbf{D}_i = 1\right] - E\left[\mathbf{Y}_{0i} \mid \mathbf{W}_{0i} = 1,\, \mathbf{D}_i = 0\right] \\
&= E\left[\mathbf{Y}_{1i} \mid \mathbf{W}_{1i} = 1\right] - E\left[\mathbf{Y}_{0i} \mid \mathbf{W}_{0i} = 1\right] \\
&= E\left[\mathbf{Y}_{1i} \mid \mathbf{W}_{1i} = 1\right] - E\left[\mathbf{Y}_{0i} \mid \mathbf{W}_{1i} = 1\right] + E\left[\mathbf{Y}_{0i} \mid \mathbf{W}_{1i} = 1\right] - E\left[\mathbf{Y}_{0i} \mid \mathbf{W}_{0i} = 1\right]
\end{aligned}
$$

Bad Controls: Occupation

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

$$
\begin{aligned}
& E\left[\mathbf{Y}_i \mid \mathbf{W}_i = 1,\, \mathbf{D}_i = 1\right] - E\left[\mathbf{Y}_i \mid \mathbf{W}_i = 1,\, \mathbf{D}_i = 0\right] \\
&= E\left[\mathbf{Y}_{1i} \mid \mathbf{W}_{1i} = 1,\, \mathbf{D}_i = 1\right] - E\left[\mathbf{Y}_{0i} \mid \mathbf{W}_{0i} = 1,\, \mathbf{D}_i = 0\right] \\
&= E\left[\mathbf{Y}_{1i} \mid \mathbf{W}_{1i} = 1\right] - E\left[\mathbf{Y}_{0i} \mid \mathbf{W}_{0i} = 1\right] \\
&= E\left[\mathbf{Y}_{1i} \mid \mathbf{W}_{1i} = 1\right] - E\left[\mathbf{Y}_{0i} \mid \mathbf{W}_{1i} = 1\right] + E\left[\mathbf{Y}_{0i} \mid \mathbf{W}_{1i} = 1\right] - E\left[\mathbf{Y}_{0i} \mid \mathbf{W}_{0i} = 1\right] \\
&= \underbrace{E\left[\mathbf{Y}_{1i} - \mathbf{Y}_{0i} \mid \mathbf{W}_{1i} = 1\right]}_{\text{ATT on white-collar workers}} + \underbrace{E\left[\mathbf{Y}_{0i} \mid \mathbf{W}_{1i} = 1\right] - E\left[\mathbf{Y}_{0i} \mid \mathbf{W}_{0i} = 1\right]}_{\text{Selection bias}}
\end{aligned}
$$

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

Bad Controls: Occupation

- Specifically,

$$\underbrace{E\left[\mathbf{Y}_{1i} - \mathbf{Y}_{0i} \mid \mathbf{W}_{1i} = 1\right]}_{\text{ATT on white-collar workers}} + \underbrace{E\left[\mathbf{Y}_{0i} \mid \mathbf{W}_{1i} = 1\right] - E\left[\mathbf{Y}_{0i} \mid \mathbf{W}_{0i} = 1\right]}_{\text{Selection bias}}$$

- The first term, $E\left[\mathbf{Y}_{1i} - \mathbf{Y}_{0i} \mid \mathbf{W}_{1i} = 1\right]$: the average causal effect of college graduation on earnings among people who would work in a white-collar job if they graduated (the ATT in this subgroup).
- If the occupational choice between white-collar and blue-collar 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 changes the wage premium between college and high school graduates.

Bad Controls v.s Omitted Variable Bias

- Bad Controls: Including bad 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: If you omit a confounder—a pretreatment variable that affects both treatment and outcome—you may suffer **OVB**, which generally yields **biased** and **inconsistent** estimates.

Bad Controls v.s Omitted Variable Bias

- Bad Controls: Including bad 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: If you omit a confounder—a pretreatment variable that affects both treatment and outcome—you may suffer **OVB**, which generally yields **biased** and **inconsistent** estimates.

Gary King, Professor at Harvard

- *"A Hard and Unsolved Problem in Social Sciences"*

  — Gary King (2010)

- It is the most "artistic" part of the econometrics in my opinion.

Wrap up

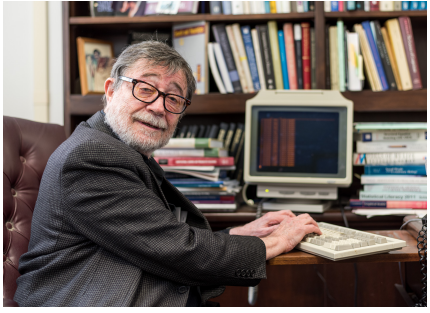
- Which variables should be included on the right-hand side of a regression equation?
  1. **Relevant and Omitted Variables**: These are variables that correlate with both the treatment and the outcome.
     - However, be careful of **bad controls**, as they can introduce more bias.
  2. **Relevant but 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**.

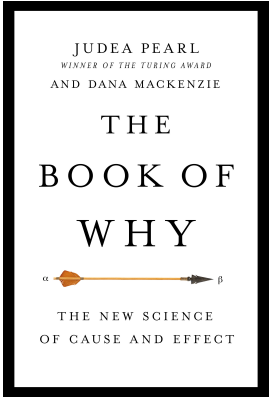
DAGs and Control Variables

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

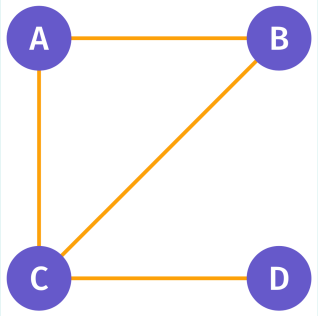




- **Computer Scientist, Turing Award Winner in 2011.**
  - *"for fundamental contributions to artificial intelligence through the development of a calculus for probabilistic and causal reasoning".*

Graphs

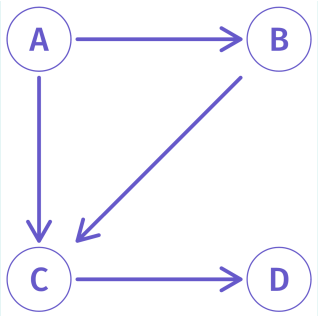
- In graph theory, a graph is a collection of **nodes** connected by **edges**



- 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

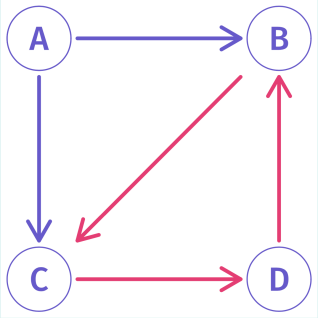
- **Directed** graphs have edges with direction: $A \to B \to C$



- **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**.

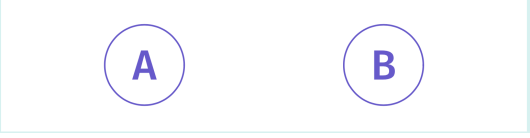


- **Directed Acyclic Graph(DAG)** is the directed graph which does not have any cycles.

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.

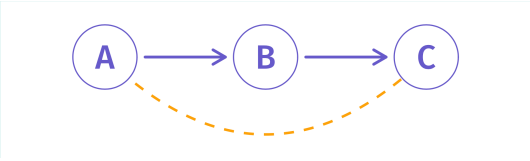
DAG: Building blocks

2. **Two connected nodes**

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

DAG: Building blocks

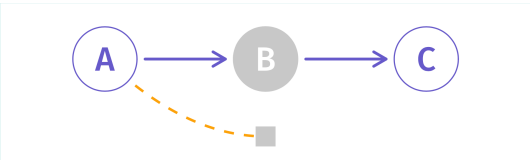
3. **Chains**



- 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

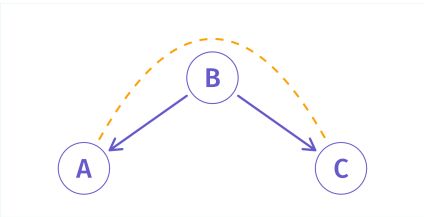
3. **Chains with conditions**



- **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

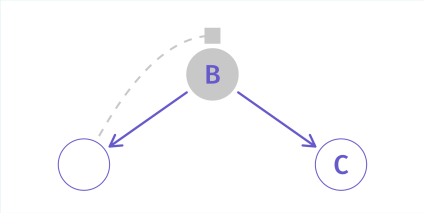
4. **Forks or Confounds**



- 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

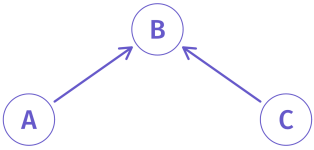
4. **Blocked forks**



- **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**



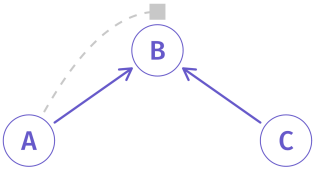
- An **immorality** occurs when two nodes share a child without being otherwise connected

$$A \rightarrow B \leftarrow C$$

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

DAG: Building blocks

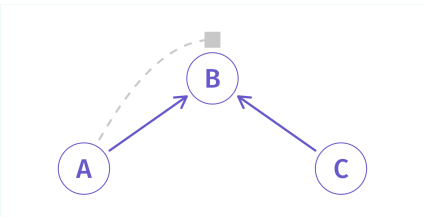
5. **Immoralities or Colliders**



- **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**

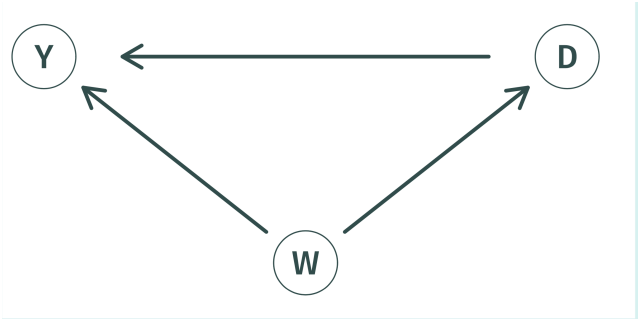


- **Question**: What happens when we condition on B?
- **Answer**: We **unblock** or **open** the previously blocked (closed) path.
  - While A and C are independent, they are conditionally dependent.
  - When you condition on a collider, you open up the path.

DAGs: Blocked Paths

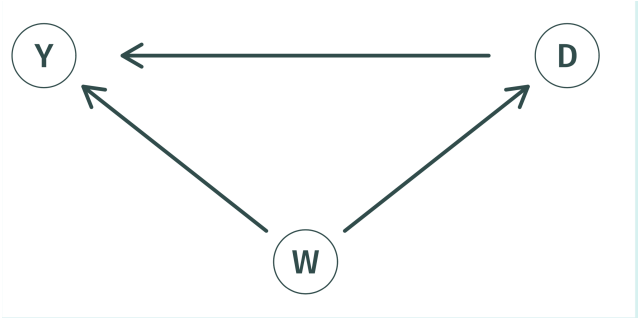
- A path between X and Y is **blocked** by conditioning on a set of variables Z if either of the following statements is true:
  1. On the path, there is a **chain** $(\cdots \to W \to \dots)$ or a **fork** $(\cdots \leftarrow W \to \dots)$, and we condition on W $(W \in Z)$.
  2. On the path, there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

OVB in a DAG



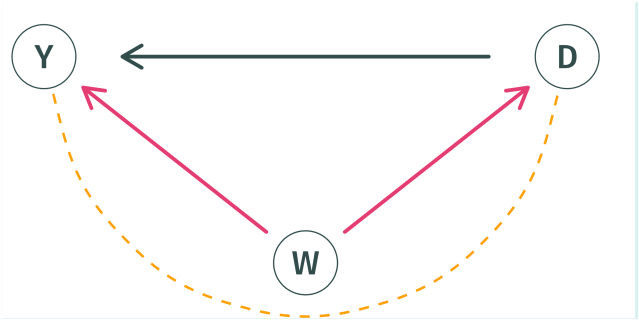
- **Nodes** are random variables.
  - Y is the dependent variable
  - D is the treatment variable
  - W is the Omitted variable

OVB in a DAG



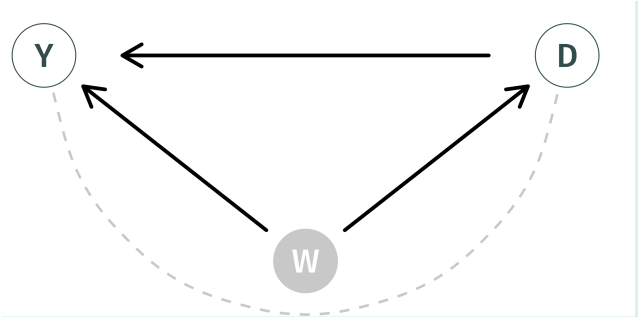
- **Edges** depict causal links.
- **Causality** flows in the direction of the arrows.
- Both connections and non-connections matter!

OVB in a DAG



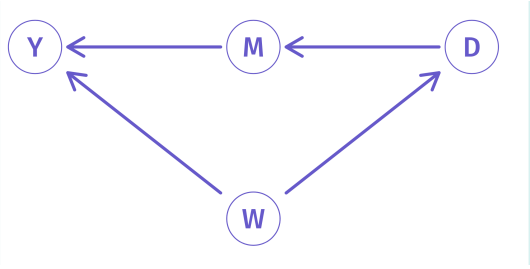
- There are two pathways from D to Y:
  1. The path from D to Y is our casual relationship of interest.
  2. The path $Y \leftarrow W \rightarrow D$ creates non-causal association between D and Y.

OVB in a DAG



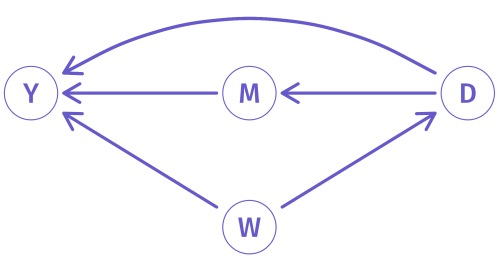
- To shut down this pathway creating a non-causal association, we have to **control/balance/adjust** on W.

Mediation in a DAG



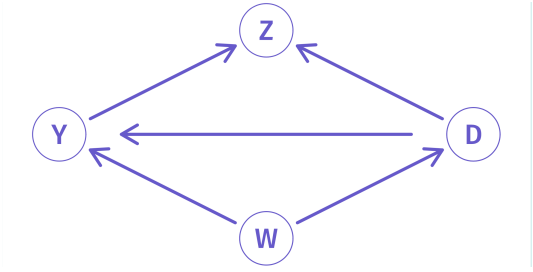
- **Question**: How to control variables to isolate the causal effect of D on Y?
- **Answer**: Control on W and Don't control on M

Partial Mediation in a DAG



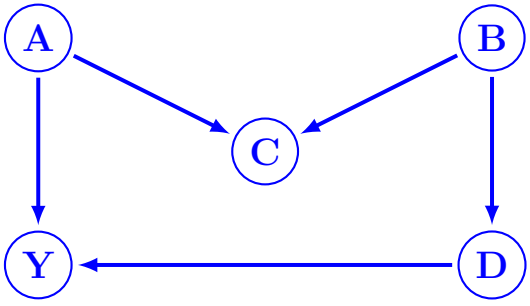
- **Question**: How to control variables to isolate the causal effect of D on Y?
- **Answer**: Control on W and Don't control on M, either.
  - Why? Because our causal effect of interest is an aggregate effect, not a partial effect.
  - Or you could underestimate the effect of D on Y.

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**:
  - Control on A which is more like a relevant variable.
  - Don't control on B which is more like a irrelevant and multicollinearity variable.

DAG Applications: Going to College on Earnings

- Our Simple OLS Regression:

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

  - 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*.

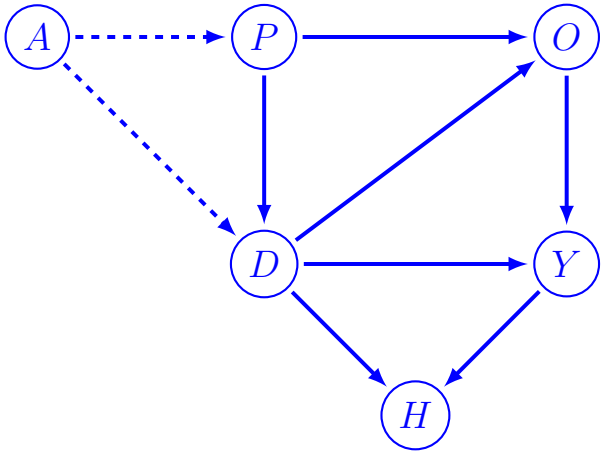
- However, the OLS estimator is biased and inconsistent because of the OVB and confounding factors.
- There are some other factors that affect the wage and schooling as well.
  - Some are observed: **Parental education(P)**, **Occupation(O)**, **Social Network(N)**, etc.
  - Some are unobserved: **Ability 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.
  - $D \rightarrow Y$ (causal relationship of interest)
  - $D \leftarrow P \rightarrow O \rightarrow Y$ (Partial Mediator)
  - $D \leftarrow P \rightarrow O \rightarrow Y$ (confounder)
  - $D \leftarrow A \rightarrow P \rightarrow O \rightarrow Y$ (confounder)
  - $D \rightarrow H \leftarrow Y$ (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**:
  - Tell which are relevant and which are not and which can be OVB
  - Control for variables that are likely to be confounders.
  - Avoid controlling for variables that might be outcomes of the treatment (post-treatment variables or collider variables) and multicollinearity variables.

Practical Tip 2

- **Sensitivity analysis**:
  - Start with a simple model which only includes the treatment variable and essential controls.
  - Add controls one by one(not too many, can be grouped) to see how they affect your estimates
  - If adding controls dramatically changes your estimate, investigate the reason why.
  - If adding controls does not change your estimate, you can be more confident about your estimate.
  - Be skeptical about “kitchen sink” regressions at first, which include every available control.

Practical Tip 3

- **Remember the fundamental trade-off**:
  - More controls can reduce omitted variable bias.
  - But they can also increase variance and potentially introduce new biases.
  - The goal is to find the “sweet spot” that balances these concerns.