# Regression and Causal Inference

Prof. Tzu-Ting Yang
楊子霆

Institute of Economics, Academia Sinica
中央研究院經濟研究所

March 20, 2023

# Regression: Main Idea

# Main Idea of Regression

- A linear regression can help us study the relationship between treatment $D_i$ and outcome $\mathrm{Y}_i$

$$\mathrm{Y}_i = \delta + \alpha D_i + \beta X_i + \epsilon_i$$

- We can interpret $\alpha$ as the causal effect of treatment by including all confounding factors $X_i$ in the regression

# Main Idea of Regression

- An example:
  - Possible source of selection bias in estimating effect of attending graduate school on earning is family wealth
    - Since people with higher family wealth might be more likely to go to graduate school
    - Also, they might have higher annual earnings no matter what they do
  - Selection bias can be eliminated by including family wealth in the regression
    - Focusing on treatment and control groups with the same family wealth

# Regression: Potential Outcomes Framework

# Conditional Independence Assumption (CIA)

## Conditional Independence Assumption

$$(\mathrm{Y}_i^1, \mathrm{Y}_i^0) \perp\!\!\!\perp D_i | X_i$$

- CIA asserts that conditional on observable characteristics $X$, potential outcomes are independent of treatment assigned $D$
  - In other words, observed covariates $X$ can fully explain the value of potential outcome for treatment and control groups
    - $\mathrm{Y}^0 = \mathrm{f}(X)$
    - $\mathrm{Y}^1 = \mathrm{f}(X)$
- CIA ensures:
  - $\mathrm{E}[\mathrm{Y}_i^0 | X_i, D_i = 1] = \mathrm{E}[\mathrm{Y}_i^0 | X_i, D_i = 0]$
  - $\mathrm{E}[\mathrm{Y}_i^1 | X_i, D_i = 1] = \mathrm{E}[\mathrm{Y}_i^1 | X_i, D_i = 0]$

# Main Idea of Regression

- We can consider regression is a special case of matching methods
  - Because both matching and regression require CIA (selection on observable) to get causal affects
  - But regression needs to assume the functional form of **potential outcomes**

# Regression and Potential Outcome

- We can run the following regression to get causal effect of $D$

$$\mathrm{Y}_i = \delta + \alpha D_i + \beta X_i + \epsilon_i$$

- It assumes that **potential outcomes** are determined by the following equations:

$$\mathrm{Y}_i^1 = \delta + \alpha + \beta X_i + \epsilon_i$$
$$\mathrm{Y}_i^0 = \delta + \beta X_i + \epsilon_i$$

- Assume $\mathrm{E}[\epsilon_i | X_i] = 0$

- $\alpha$ is the causal effect of treatment
  - $\mathrm{E}[\mathrm{Y}_i^1 - \mathrm{Y}_i^0 | X_i] = \alpha$

# Identification Results for Regression

- Based on CIA, including key observed covariates $X_i$ into regression can help us eliminate selection bias
- Therefore, the estimated coefficient of treatment $D$ is the following:

$$\alpha(X) = \underbrace{\mathrm{E}[\mathrm{Y}_i|X_i, D_i = 1] - \mathrm{E}[\mathrm{Y}_i|X_i, D_i = 0]}_{\text{Observed Difference in Average Outcome at given } X_i}$$

# Identification Results for Regression

$$\alpha(X) = \underbrace{\mathrm{E}[\mathrm{Y}_i|X_i, D_i = 1] - \mathrm{E}[\mathrm{Y}_i|X_i, D_i = 0]}_{\text{Observed Difference in Average Outcome at given } X_i}$$

$$= \mathrm{E}[\mathrm{Y}_i^1|X_i, D_i = 1] - \mathrm{E}[\mathrm{Y}_i^0|X_i, D_i = 1]$$

$$+ \mathrm{E}[\mathrm{Y}_i^0|X_i, D_i = 1] - \mathrm{E}[\mathrm{Y}_i^0|X_i, D_i = 0]$$

$$= \underbrace{\mathrm{E}[\mathrm{Y}_i^1 - \mathrm{Y}_i^0|X_i, D_i = 1]}_{\text{Causal Effect (CATT)}}$$

$$+ \underbrace{\mathrm{E}[\mathrm{Y}_i^0|X_i, D_i = 1] - \mathrm{E}[\mathrm{Y}_i^0|X_i, D_i = 0]}_{\text{Selection Bias}}$$

$$= \underbrace{\mathrm{E}[\mathrm{Y}_i^1 - \mathrm{Y}_i^0|X_i, D_i = 1]}_{\text{Causal Effect (CATT)}}$$

$$+ \underbrace{\beta\mathrm{E}[\mathrm{X}_i|X_i, D_i = 1] - \beta\mathrm{E}[\mathrm{X}_i|X_i, D_i = 0]}_{\text{Selection Bias}}$$

# Identification Results for Regression

$$
\begin{aligned}
\alpha(X) = \quad & \underbrace{\mathrm{E}[\mathrm{Y}_i|X_i, D_i = 1] - \mathrm{E}[\mathrm{Y}_i|X_i, D_i = 0]}_{\text{Observed Difference in Average Outcome at given } X_i} \\
&= \underbrace{\mathrm{E}[\mathrm{Y}_i^1 - \mathrm{Y}_i^0|X_i, D_i = 1]}_{\text{Causal Effect (CATT)}} \\
&+ \underbrace{\beta\mathrm{E}[\mathrm{X}_i|X_i, D_i = 1] - \beta\mathrm{E}[\mathrm{X}_i|X_i, D_i = 0]}_{\text{Selection Bias}} \\
&= \underbrace{\mathrm{E}[\mathrm{Y}_i^1 - \mathrm{Y}_i^0|X_i, D_i = 1]}_{\text{Causal Effect (CATT)}} + \underbrace{0}_{\text{Selection Bias}} \\
&= \underbrace{\mathrm{E}[\mathrm{Y}_i^1 - \mathrm{Y}_i^0|X_i, D_i = 0]}_{\text{Causal Effect (CATU)}} \\
&= \underbrace{\mathrm{E}[\mathrm{Y}_i^1 - \mathrm{Y}_i^0|X_i]}_{\text{Causal Effect (CATE)}}
\end{aligned}
$$

# Identification Results for Regression

- Note that there are as many causal effects (CATE or CATT) as the number of value in $X_i$
- People might find it useful to boil a set of estimates down to a single summary measure
  - e.g. Average treatment effect
- Again, applying the law of iterated expectations (LIE), we can identify ATT, ATU, and ATE
  - Take average of CATT, CATU, and CATE over all subgroups (all possible X-values)

# Regression: Estimation

# Regression: Estimation

- Again, if we have population data, we can get the above causal effect $\alpha$
- However, we usually only have sample (i.e. part of population data)
- We need to use sample data to estimate $\alpha$

# Review: Ordinary Least Squares Estimation

- Regression analysis assigns values to model parameters ($\delta$ and $\alpha$) so as to make $\hat{\mathrm{Y}}_i$ as close as possible to $\mathrm{Y}_i$

- Ordinary Least Squares (OLS) estimation accomplish it by choosing values that minimize the sum of squared residuals

$$(\hat{\delta}, \hat{\alpha}) = \min_{\delta,\alpha} \frac{1}{N} \sum_{i=1}^{N} (\mathrm{Y}_i - \delta - \alpha D_i)^2$$

- OLS estimator for treatment effect $\alpha$:

$$\hat{\alpha} = \frac{Cov(\mathrm{Y}_i, D_i)}{V(D_i)}$$

# Review: Omitted Variable Bias

- Failure to include enough (right) control variables in the regression would result in selection bias
- The **OLS version** of the **selection bias** generated by inadequate controls is called **Omitted Variable Bias (OVB)**

# Review: Omitted Variable Bias

- Suppose the true model is:

$$\mathrm{Y}_i = \delta + \alpha D_i + \beta X_i + \epsilon_i$$

- $X_i$ is the observed characteristics (e.g. family wealth)
- But we estimate this model:

$$\mathrm{Y}_i = \delta + \alpha D_i + u_i$$

- where $u_i = \beta X_i + \epsilon_i$
- Assume $\mathrm{E}[\epsilon_i | X_i] = 0$

# Review: Omitted Variable Bias

- OVB formula:

$$\hat{\alpha} \xrightarrow{p} \alpha + \frac{Cov(u_i, D_i)}{V(D_i)}$$
$$= \alpha + \beta \frac{Cov(X_i, D_i)}{V(D_i)}$$

- The difference between estimated treatment effect $\hat{\alpha}$ and true effect $\alpha$ depends on two components:

  1. $\beta$: The effect of omitted variable $X_i$ on outcome variable $\mathrm{Y}_i$
  2. $\frac{Cov(X_i, D_i)}{V(D_i)}$: The relationship between omitted variable $X_i$ and treatment variable $D_i$

# Review: Omitted Variable Bias

$$\begin{tikzcd}
 & X \arrow[dl] \arrow[dr] & \\
D & & Y
\end{tikzcd}$$

- The confounding factor $X$ can result in the co-movement between treatment $D$ and outcome $Y$
- Even if treatment $D$ has no causal effect on outcome $Y$

# Review: Omitted Variable Bias

## Example

- OVB formula:

$$\hat{\alpha} \xrightarrow{p} \alpha + \frac{Cov(u_i, D_i)}{V(D_i)}$$
$$= \alpha + \beta \frac{Cov(X_i, D_i)}{V(D_i)}$$

- The difference between estimated effect of attending graduate school $\hat{\alpha}$ and true effect of attending graduate school $\alpha$ depends on two components:

1. $\beta$: The effect of family wealth (omitted) $X_i$ on earnings $\mathrm{Y}_i$
2. $\frac{Cov(X_i, D_i)}{V(D_i)}$: The relationship between family wealth $X_i$ and attending graduate school $D_i$

# Review: Omitted Variable Bias

- In RCT, we can eliminate OVB since treatment assignment $D_i$ is unrelated to other confounding factors $X_i$
  - $\frac{Cov(X_i, D_i)}{V(D_i)} = 0$
- In the regression, we can eliminate OVB by including other **observed** confounding factors $X_i$ into regression
  - $\frac{Cov(u_i, D_i)}{V(D_i)} = 0$
  - When we include $X_i$ in regression model, $u_i = \epsilon_i$ which is unrelated to treatment status $D_i$

# Review: Omitted Variable Bias

- OVB formula is a tool that allows us to consider the impact of controlling for variables we wish we had
- We cannot use data to check the consequences of omitted variables that we do not observe
- But we can use the OVB formula to make a educated guess as to the likely consequences of their omission

$$\hat{\alpha} \xrightarrow{p} \alpha + \beta \frac{Cov(X_i, D_i)}{V(D_i)}$$

# Summary of Hypothesis Testing for Regression

- We estimate the following regression and want to test whether there is treatment effect:

$$Y_i = \delta + \alpha D_i + \beta X_i + \epsilon_i$$

1. Choose a null hypothesis:
    - We usually test whether there is **no average effect** of treatment
    - $H_0 : \alpha = 0$

# Summary of Hypothesis Testing for Regression

2. Choose a test statistic

   - We use a t-statistic to measure whether our sample estimates support/against this null hypothesis
   - $t = \dfrac{(\hat{\alpha} - \alpha)}{\hat{\mathrm{SE}}(\hat{\alpha})}$

# Summary of Hypothesis Testing for Regression

3. Estimate standard error of the estimator

- $\hat{\mathrm{SE}}(\hat{\alpha}) = \sqrt{\dfrac{V(\overline{D}_i \epsilon_i)}{N_s(\hat{\sigma_{\overline{D}}})^4}}$
  - $\hat{\sigma_{\overline{D}}}$ is the standard deviation of $\overline{D_i}$
  - $\overline{D_i}$ is the residual from a regression of $D_i$ on all other regressors $X_i$
- The addition of covariates $X$ has two opposing effects on $\hat{\mathrm{SE}}(\hat{\alpha})$.
  1. $\hat{\sigma_{\overline{D}}}$ might decrease since addition covariates explain some of the variation in other regressors
  2. The residual variance $V(\overline{D_i}\epsilon_i)$ falls when covariates that predict $Y_i$ are added to the regressions
- This is known as **robust standard errors**

# Summary of Hypothesis Testing for Regression

4. Evaluate whether the sample estimator is against null hypothesis or not

   - **Goal:** Calculate **p-value**

     - **p-value:** Given null hypothesis is true, the probability of obtaining the sample estimates or more extreme ones

     - If this probability is high, it means the sample estimate might support for null hypothesis

     - If this probability is low, it means the sample estimate might be against null hypothesis

# Summary of Hypothesis Testing for Regression

4. Evaluate whether the sample estimator is against null hypothesis or not

   - In order to calculate this probability (p-value), we need to know the distribution of the t-statistic under the null hypothesis

     - If sample size is sufficiently large, using **Central Limit Theorem (CLT)**, t-statistic will have standard normal distribution

# Summary of Hypothesis Testing for Regression

4. Evaluate whether the sample estimator is against null hypothesis or not

    - Based on standard normal distribution and sample estimator, we can get p-value
    - We reject the null hypothesis $H_0 : \alpha = 0$ when p-value is sufficiently low
        - We usually select an arbitrarily pre-defined threshold value $\theta$, which is referred to as the **level of significance**
        - By convention, $\theta$ is commonly set to 0.1 or 0.05
    - If p-value is smaller than $\theta$, we would say the sample estimate is **significantly different from the null hypothesis**

# Interpretation of Regression Results

- Suppose the estimated regression is the following:

$$\hat{\mathrm{Y}}_{\mathrm{i}} = 35000 + 5000 D_i + 0.5 X_i$$

- Suppose the estimated standard error is:

$$\hat{\mathrm{SE}}(\hat{\alpha}) = 1000$$

- So the t-statistic for testing $H_0 : \alpha = 0$:

$$t = \frac{(\hat{\alpha} - \alpha)}{\hat{\mathrm{SE}}(\hat{\alpha})} = \frac{5000 - 0}{1000} = 5$$

# Interpretation of Regression Results

- Using t-statistic, we can compute the p-value $= 0.00001$, which is much lower than 0.05 or 0.01
  - Given null hypothesis $H_0 : \alpha = 0$ is true, our estimate is unlikely to happen (but it happens!!)
  - It suggests our estimate is against the null hypothesis
  - Thus, we should reject the null hypothesis

# Interpretation of Regression Results

- Based on sample estimates and its standard deviation, we can construct a confidence interval for $\alpha$

- Note that the t-statistic for 5% two-sided significance level is 1.96

$$\hat{\alpha} \pm 1.96\hat{\mathrm{SE}}(\hat{\alpha}) = 5000 \pm 1.96 \times 1000$$

  - The 95% confidence interval does not include zero

  - Null hypothesis $H_0 : \alpha = 0$ is rejected at the 5% level

# STATA Command: reg

- **reg**: Linear regression
- Syntax:

```
reg depvar [indepvars] [if] [in] [weight] [,options]
```

# STATA Command: reg

- Please see **reg.do**

```
reg incwage college i.health age year i.race, vce(robust)
predict incwage_hat
predict incwage_hat_std, stdp
```

- To examine the effect of college degree on wage by controlling many demographic factors

- Option **vce(robust)**: use robust standard error

- **predict**: creates newvar containing linear prediction xb for whole sample

- Option **stdp**: creates newvar containing the standard error of the linear prediction xb

# STATA Command: reg

## Output

```
. reg incwage college i.health age year i.race, vce(robust)

Linear regression                               Number of obs     =     46,299
                                                F(22, 46275)      =          .
                                                Prob > F          =          .
                                                R-squared         =     0.1106
                                                Root MSE          =      48789

------------------------------------------------------------------------------
             |               Robust
     incwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     college |   32653.63   658.7384    49.57   0.000     31362.49    33944.76
             |
      health |
  very good  |  -1385.204   661.5773    -2.09   0.036    -2681.905   -88.50227
       good  |  -7219.177   672.5077   -10.73   0.000    -8537.302   -5901.052
       fair  |  -17981.89   775.4887   -23.19   0.000    -19501.86   -16461.92
       poor  |  -25661.95   771.2066   -33.28   0.000    -27173.53   -24150.37
```

# STATA Command: reg

## Output

| | incwage | incwage_hat | incwage_ha~d |
|---|---|---|---|
| 1 | 15000 | 53263.65 | 858.3494 |
| 2 | 42000 | 53174.29 | 854.6627 |
| 3 | 29000 | 49778.7 | 813.7073 |
| 4 | . | 22020.95 | 561.3401 |
| 5 | . | 21484.8 | 575.526 |

# STATA Command: reg

```
reg incwage college i.health age year i.race if sex
    ==1, vce(robust)
predict incwage_hat_m if e(sample)
```

- Option **if**: restrict sample to specific subgroup
- Option **if e(sample)**: obtain linear prediction for male (if $sex == 1$)

# STATA Command: reg

## Output

```
. reg incwage college i.health age year i.race if sex==1, vce(robust)

Linear regression                               Number of obs     =     22,173
                                                F(19, 22149)      =          .
                                                Prob > F          =          .
                                                R-squared         =     0.1303
                                                Root MSE          =      58415
```

| incwage | Coef. | Robust Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| college | 43120.23 | 1206.212 | 35.75 | 0.000 | 40755.96 | 45484.49 |
| health | | | | | | |
| very good | -2554.815 | 1126.348 | -2.27 | 0.023 | -4762.537 | -347.0923 |
| good | -9808.018 | 1160.95 | -8.45 | 0.000 | -12083.56 | -7532.473 |
| fair | -23376.74 | 1418.75 | -16.48 | 0.000 | -26157.59 | -20595.89 |
| poor | -34504.03 | 1383.097 | -24.95 | 0.000 | -37215 | -31793.06 |

# Suggested Readings

- Chapter 2, Mastering Metrics: The Path from Cause to Effect
- Chapter 3, Mostly Harmless Econometrics
- Chapter 2, Causal Inference: The Mixtape