Regression 1: Framework

Instructor: Yuta Toyama

Last updated: 2021-06-16

Introduction

Observational Study (観察研究)

- Researchers in social science cannot always conduct a randomized control trial.
- Instead, we need to use **observational data** in which treatment assignment may not be random.
- An approach in this case is **controlling observable characteristics** that causes a selection bias.
- This approach is essentially estimation of **linear regression model (線形回帰モデル)** by **ordinally least squares (OLS, 最小二乗法)**.

Overview

- Introduce an idea of **matching (マッチング)** estimator.
  - Identification of treatment effect under **selection on observable** assumption.
  - Linear regression is a special case of matching estimator.
- Linear regression: framework, practical topics, inference

Selection on Observables, or Matching

Matching to eliminate a selection bias

- Idea: Compare **individuals with the same observed characteristics** X across treatment and control groups
- If treatment choice is driven by observed characteristics (such as age, income, gender, etc), controlling for such factor would eliminate the selection.
- Two key assumptions in matching

Assumption 1: Selection on observables

- Let X_i denote the observed characteristics (sometimes called **covariates (共変量)**)
  - age, income, education, race, etc..
- Assumption 1:

$$D_i \perp (Y_{0i}, Y_{1i}) \,|X_i$$

- **Conditional on X_i**, treatment assignment is random.
- This assumption is often referred in a different name:
  - Selection on observables
  - Ignorability
  - Unconfoundedness

Assumption 2: Overlapping assumption

- Assumption 2:

$$P(D_i = 1 | X_i = x) \in (0, 1) \; \forall x$$

- Given x, we should be able to observe people from both control and treatment group.
- The probability $P(D_i = 1 | X_i = x)$ is called **propensity score (傾向スコア)**.

Identification of Treatment Effect Parameters

- Assumption 1 (unconfoundedness) implies that

$$
\begin{aligned}
E[Y_{1i}|D_i = 1, X_i] = E[Y_{1i}|D_i = 0, X_i] = E[Y_{1i}|X_i] \\
E[Y_{0i}|D_i = 1, X_i] = E[Y_{0i}|D_i = 0, X_i] = E[Y_{0i}|X_i]
\end{aligned}
$$

- Once you conditioning on X_i, the argument is essentially the same as the one in RCT.

- The ATT conditional on $X_i = x$ is given by

$$\begin{aligned} E[Y_{1i} - Y_{0i}|D_i = 1, X_i] &= E[Y_{1i}|D_i = 1, X_i] - E[Y_{0i}|D_i = 1, X_i] \\ &= E[Y_{1i}|D_i = 1, X_i] - E[Y_{0i}|D_i = 0, X_i] \end{aligned}$$

- Assumption 2 (overlapping) is needed to use the following

$$E[Y_{di}|D_i = d, X_i] = E[Y_i|D_i = d, X_i] \text{ for } d = 0, 1$$

- Why? Overlapping assumption $P(D_i = 1|X_i = x) \in (0, 1)$ means that for each x, **we should have people in both treatment and control group**.

- If not, we cannot observe both $E[Y_i|D_i = d, X_i]$ for $d = 0, 1$.

- With two assumptions,

$$E[Y_{1i} - Y_{0i}|D_i = 1, X_i] = \underbrace{E[Y_i|D_i = 1, X_i]}_{avg\ with\ X_i\ in\ treatment} - \underbrace{E[Y_i|D_i = 0, X_i]}_{avg\ with\ X_i\ in\ control}$$

ATT $E[Y_{1i} - Y_{0i}|D_i = 1]$

- ATT is given by

$$\begin{aligned} ATT &= E[Y_{1i} - Y_{0i}|D_i = 1] \\ &= \int E[Y_{1i} - Y_{0i}|D_i = 1, X_i = x] f_{X_i}(x|D_i = 1)dx \\ &= E[Y_i|D_i = 1] - \int \left(E[Y_i|D_i = 0, X_i = x]\right) f_{X_i}(x|D_i = 1) \end{aligned}$$

ATT $E[Y_{1i} - Y_{0i}]$

- ATE is

$$
\begin{aligned}
ATE =& E[Y_{1i} - Y_{0i}] \\
=& \int E[Y_{1i} - Y_{0i} | X_i = x] f_{X_i}(x) dx \\
=& \int E[Y_{1i} | D_i = 1, X_i = x] f_{X_i}(x) dx + \int E[Y_{0i} | D_i = 0, X_i = x] f_{X_i}(x) dx \\
=& \int E[Y_i | D_i = 1, X_i = x] f_{X_i}(x) dx + \int E[Y_i | D_i = 0, X_i = x] f_{X_i}(x) dx
\end{aligned}
$$

From Identification to Estimation

- We need to estimate two conditional expectations $E[Y_i|D_i = 1, X_i = x]$ and $E[Y_i|D_i = 0, X_i = x]$
- Several ways to implement this.
    1. Regression: Nonparametric and Parametric
    2. Nearest neighborhood matching (最近傍マッチング)
    3. Propensity Score Matching (傾向スコアマッチング)
- Here, I only explain a parametric regression as a way to implement the matching method.
- See Appendix and textbooks for the details of matching estimators.

From Matching to Linear Regression Model

- Assume that

$$E[Y_i|D_i = 0, X_i = x] = \beta' x_i$$
$$E[Y_i|D_i = 1, X_i = x] = \beta' x_i + \tau$$

- Here, treament effect is given by τ.
- You will have a linear regression model

$$y_i = \beta' x_i + \tau D_i + \epsilon_i, E[\epsilon_i|D_i, x_i] = 0$$

- Running a linear regression to obtain the treatment effect parameter τ.

Linear Regression: Framework

Regression (回帰) framework

- **Linear regression model (線形回帰モデル)** is defined as

$$Y_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_K X_{Ki} + \epsilon_i$$

  - i: index for observations. $i = 1, \cdots, N$.
  - Y_i: **dependent variable (被説明変数)**
  - X_{ki}: **explanatory variable (説明変数)**
  - ϵ_i: **error term (誤差項)**
  - β: **coefficients (係数)**

- Data (sample): $\{Y_i, X_{i1}, \ldots, X_{iK}\}_{i=1}^{N}$

- We want to estimate coefficients β.

Ordinaly Least Squares (最小二乗法、OLS)

- OLS estimators are the minimizers of the sum of squared residuals:

$$\min_{\beta_0,\cdots,\beta_K} \frac{1}{N} \sum_{i=1}^{N} (Y_i - (\beta_0 + \beta_1 X_{i1} + \cdots + \beta_K X_{iK}))^2$$

- First order conditions characterize the OLS estimator. Denote it by $\hat{\beta}$.

Residual Regression (残差回帰)

- Consider the model

$$Y_i = \beta_0 + \alpha D_i + \beta_1 X_{1i} + \cdots + \beta_K X_{Ki} + \epsilon_i$$

- Suppose that you are interested in α (say treatment effect parameter).
- Residual regression characterizes the OLS estimator of $\hat{\alpha}$ in the following way.

Frisch–Waugh–Lovell Theorem

1. Run OLS regression of D_i on all other explanatory variables $1, X_{1i}, \cdots, X_{Ki}$. Obtain the residual $\hat{u_i^D}$.

2. Run OLS regression of Y_i on all other explanatory variables $1, X_{1i}, \cdots, X_{Ki}$. Obtain the residual $\hat{u_i^Y}$.

3. Run OLS regression of $\hat{u}_i^Y$ on $\hat{u}_i^D$ without constant term. The OLS estimator $\hat{\alpha}$ is

$$\hat{\alpha} = \frac{\sum \hat{u}_i^Y \hat{u}_i^D}{\sum (\hat{u}_i^D)^2}$$

How to use FWL theorem

1. Computational advantage if you are interested in a particular coefficient. Use this idea in estimation of panel data model.

1. Useful to see how the coefficient of interest is estimated. We will see this later in relation to multicolinearity (多重共線性).

1. Double machine learning (Chernozhukov et al 2018): Estimation of treatment effect parameters when so many covariates are available.

Assumptions for OLS

1. **Random sample (ランダムサンプル)**: $\{Y_i, X_{i1}, \ldots, X_{iK}\}$ is i.i.d. (identically and independently distributed) drawn sample
2. **mean independence**: ϵ_i has zero conditional mean
$$E[\epsilon_i | X_{i1}, \ldots, X_{iK}] = 0$$
3. Large outliers are unlikely: The random variable Y_i and X_{ik} have finite fourth moments.
4. **No perfect multicollinearity (多重共線性)**: No linear relationship between explanatory variables.

Theoretical Properties of OLS estimator

1. **Unbiasedness**: Conditional on the explantory variables X, the expectation of the OLS estimator $\hat{\beta}$ is equal to the true value β.

$$E[\hat{\beta}|X] = \beta$$

2. **Consistency**: As the sample size N goes to infinity, the OLS estimator $\hat{\beta}$ converges to β in probability

$$\hat{\beta} \xrightarrow{p} \beta$$

3. **Asymptotic normality (渐近正规性)**: discuss later.

Linear Regression: Practical Topics

Interpretation of Regression Coefficients

- Remember that

$$Y_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_K X_{Ki} + \epsilon_i$$

- The coefficient β_k: the effect of X_k on Y **ceteris paribus (all things being equal)**
- Equivalently, if X_k is continuous random variable,

$$\frac{\partial Y}{\partial X_k} = \beta_k$$

- If we can estimate β_k without bias, can obtain **causal effect** of X_k on Y.

Common Specifications in Linear Regression Model

- Several specifications frequently used in empirical analysis.
  1. Nonlinear term
  2. log specification
  3. dummy (categorical) variables
  4. interaction terms (交差項)

Nonlinear term (非線形項)

- Non-linear relationship between Y and X in a linearly additive form

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 X_i^3 + \epsilon_i$$

- As long as the error term ϵ_i appreas in a additively linear way, we can estimate the coefficients by OLS.
    - Multicollinarity could be an issue if we have many polynomials (多項式).
    - You can use other non-linear variables such as $\log(x)$ and $\sqrt{x}$.

log specification

- Using `log` changes the interpretation of the coefficient β in terms of scales.

Dependent	**Explanatory**	**interpretation**
Y	X	1 unit increase in X causes β units change in Y
$\log Y$	X	1 unit increase in X causes $100\beta\%$ change in Y
Y	$\log X$	1% increase in X causes $\beta/100$ unit change in Y
$\log Y$	$\log X$	1% increase in X causes $\beta\%$ change in Y

Dummy variable (ダミー変数)

- A **dummy variable** takes only 1 or 0. This is used to express qualititative information
- Example: Dummy variable for race

$$white_i = \begin{cases} 1 & if\, white \\ 0 & otherwise \end{cases}$$

- The coefficient on a dummy variable captures the difference of the outcome Y between categories

$$Y_i = \beta_0 + \beta_1 white_i + \epsilon_i$$

The coefficient β_1 captures the difference of Y between white and non-white people.

Interaction term (交差項)

- You can add the interaction of two explanatory variables in the regression model.
- For example:

$$wage_i = \beta_0 + \beta_1 educ_i + \beta_2 white_i + \beta_3 educ_i \times white_i + \epsilon_i$$

  where $wage_i$ is the earnings of person i and $educ_i$ is the years of schooling for person i.
- The effect of $educ_i$ is

$$\frac{\partial wage_i}{\partial educ_i} = \beta_1 + \beta_3 white_i,$$

- This allows for heterogeneous effects of education across races.

Measures of Fit

- We often use R^2 **(決定係数)** as a measure of the model fit.
- Denote **the fitted value** as $\hat{y}_i$

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \cdots + \hat{\beta}_K X_{iK}$$

  - Also called prediction from the OLS regression.

- R^2 is defined as

$$R^2 = \frac{SSE}{TSS},$$

  where

$$SSE = \sum_i (\hat{y}_i - \bar{y})^2, \; TSS = \sum_i (y_i - \bar{y})^2$$

- R^2 captures the fraction of the variation of Y explained by the regression model.
- Adding variables always (weakly) increases R^2.

- In a regression model with multiple explanatory variables, we often use **adjusted** R^2 that adjusts the number of explanatory variables

$$\bar{R}^2 = 1 - \frac{N-1}{N-(K+1)} \frac{SSR}{TSS}$$

where

$$SSR = \sum_i (\hat{y}_i - y_i)^2 (= \sum_i \hat{u}_i^2)$$

Linear Regression: Inference

Statistical Inference of OLS Estimator

- The OLS estimator is **random variables** as it depends on a drawn sample.
- We need to conduct **statistical inference** to evaluate statistical uncertainty of the OLS estimates.
- Plan
  - Asymptotic distribution (漸近分布) of OLS estimator
  - Statistical inference:
  - Homoskedasticity (均一分散) vs Heteroskedasticity (不均一分散)

Asymptotic Normality (漸近正規性) of OLS Estimator

- Under the OLS assumption, the OLS estimator has **asymptotic normality**

$$\sqrt{N}(\hat{\beta}-\beta) \xrightarrow{d} N(0, V)$$

- V is called **asymptotic variance (matrix)** given by

$$\underbrace{V}_{(K+1)\times(K+1)} = E[\mathbf{x}_i'\mathbf{x}_i]^{-1}E[\mathbf{x}_i'\mathbf{x}_i\epsilon_i^2]E[\mathbf{x}_i'\mathbf{x}_i]^{-1}$$

- $\mathbf{x}_i = (1, X_{i1}, \cdots, X_{iK})'$ is $(K+1)\times 1$ vector.

- We can **approximate** the distribution of $\hat{\beta}$ by

$$\hat{\beta} \sim N(\beta, V/N)$$

- The individual coefficient β_k follows

$$\hat{\beta}_k \sim N(\beta_k, V_{kk}/N)$$

Estimation of Asymptotic Variance (漸近分散)

- V is an unknown object. Need to be estimated.
- Consider the estimator $\hat{V}$ for V using sample analogues

$$\hat{V} = \left(\frac{1}{N} \sum_{i=1}^{N} \mathbf{x}_i' \mathbf{x}_i \right)^{-1} \left(\frac{1}{N} \sum_{i=1}^{N} \mathbf{x}_i' \mathbf{x}_i \hat{\epsilon}_i^2 \right) \left(\frac{1}{N} \sum_{i=1}^{N} \mathbf{x}_i' \mathbf{x}_i \right)^{-1}$$

  where $\hat{\epsilon}_i = y_i - (\hat{\beta}_0 + \cdots + \hat{\beta}_K X_{iK})$ is the residual.
- Technically speaking, $\hat{V}$ converges to V in probability.
- We often use the (asymptotic) **standard error** $SE(\hat{\beta}_k) = \sqrt{\hat{V}_{kk}/N}$.
- The standard error is an estimator for the standard deviation of the OLS estimator $\hat{\beta}_k$.

Hypothesis testing

- You might want to test a particular hypothesis regarding those coefficients.
  - Does x really affects y?
  - Is the production technology the constant returns to scale?

3 Steps in Hypothesis Testing

- Step 1: Consider the null hypothesis H_0 and the alternative hypothesis H_1

$$H_0 : \beta_1 = k, H_1 : \beta_1 \neq k$$

  where k is the known number you set by yourself.

- Step 2: Define **t-statistic** by

$$t_n = \frac{\hat{\beta}_1 - k}{SE(\hat{\beta}_1)}$$

- Step 3: We reject H_0 is at α-percent significance level if

$$|t_n| > C_{\alpha/2}$$

  where $C_{\alpha/2}$ is the $\alpha/2$ percentile of the standard normal distribution. We say we **fail to reject** H_0 if the above does not hold.

Caveats on Hypothesis Testing

- We often say $\hat{\beta}$ is **statistically significant (統計的有意)** at 5% level if $|t_n| > 1.96$ when we set $k = 0$.

- You should also discuss **economic significance (経済的有意)** of the coefficient in analysis.

- Case 1: Small but statistically significant coefficient.
  - As the sample size N gets large, the SE decreases.

- Case 2: Large but statistically insignificant coefficient.
  - The variable might have an important (economically meaningful) effect.
  - But you may not be able to estimate the effect precisely with the sample at your hand.

F test

- We often test a composite hypothesis that involves multiple parameters such as

$$H_0 : \beta_1 + \beta_2 = 0,\ H_1 : \beta_1 + \beta_2 \neq 0$$

- We use **F test** in such a case.

Confidence interval (信頼区間)

- 95% confidence interval

$$CI_n = \left\{ k : |\frac{\hat{\beta}_1 - k}{SE(\hat{\beta}_1)}| \leq 1.96 \right\}$$
$$= \left[\hat{\beta}_1 - 1.96 \times SE(\hat{\beta}_1), \hat{\beta}_1 + 1.96 \times SE(\hat{\beta}_1) \right]$$

- Interpretation: If you draw many samples (dataset) and construct the 95% CI for each sample, 95% of those CIs will include the true parameter.

Homoskedasticity vs Heteroskedasticity

- The error term ϵ_i has **heteroskedasticity (不均一分散)** if $Var(u_i|X_i)$ depends on X_i. The asymptotic variance is

$$V = E[\mathbf{x}_i'\mathbf{x}_i]^{-1}E[\mathbf{x}_i'\mathbf{x}_i\epsilon_i^2]E[\mathbf{x}_i'\mathbf{x}_i]^{-1}$$

- If not, we call ϵ_i has **homoskedasticity (均一分散)**. In this case,

$$V = E[\mathbf{x}_i'\mathbf{x}_i]^{-1}\sigma^2$$

  where $\sigma^2 = V(\epsilon_i)$.

Standard Errors in Practice

- Standard errors under heteroskedasticity assumption is called **heteroskedasticity robust standard errors (不均一分散に頑健な標準誤差)**
- In many statistical packages (including R and Stata), the standard errors for the OLS estimators are calculated under homoskedasticity assumption as a default.
- However, if the error has heteroskedasticity, the standard error under homoskedasticity assumption will be **underestimated**.
- In OLS, **we should always use heteroskedasticity robust standard error.**

Appendix: Matching Estimator

Estimation Methods

- We need to estimate $E[Y_i|D_i = 1, X_i = x]$ and $E[Y_i|D_i = 0, X_i = x]$
- Several ways to implement the above idea
    1. Regression: Nonparametric and Parametric
    2. Nearest neighborhood matching
    3. Propensity Score Matching

Approach 1: Regression, or Analogue Approach

- Let $\hat{\mu}_k(x)$ be an estimator of $\mu_k(x) = E[Y_i | D_i = k, X_i = x]$ for $k \in \{0, 1\}$
- The analog estimators are

$$A\hat{T}E = \frac{1}{N} \sum_{i=1}^{N} (\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i))$$

$$A\hat{T}T = \frac{N^{-1} \sum_{i=1}^{N} D_i(Y_i - \hat{\mu}_0(X_i))}{N^{-1} \sum_{i=1}^{N} D_i}$$

- How to estimate $\mu_k(x) = E[Y_i | D_i = k, X_i = x]$?

Nonparametric Estimation

- Suppose that $X_i \in \{x_1, \cdots, x_K\}$ is discrete with small K
  - Ex: two demographic characteristics (male/female, white/non-white). $K = 4$ \bigskip
- Then, a nonparametric binning estimator is

$$\hat{\mu}_k(x) = \frac{\sum_{i=1}^{N} \mathbf{1}\{D_i = k, X_i = x\} Y_i}{\sum_{i=1}^{N} \mathbf{1}\{D_i = k, X_i = x\}}$$

  \bigskip
- Here, I do not put any parametric assumption on $\mu_k(x) = E[Y_i | D_i = k, X_i = x]$.

Curse of dimensionality

- Issue: Poor performance if K is large due to many covariates.
  - So many potential groups, too few observations for each group.
  - With K variables, each of which takes L values, L^K possible groups (bins) in total.
- This is known as **curse of dimensionality**.
- Relatedly, if X is a continuous random variable, can use kernel regression.

Parametric Estimation, or going back to linear regression

- If you put parametric assumption such as

$$E[Y_i|D_i = 0, X_i = x] = \beta' x_i$$
$$E[Y_i|D_i = 1, X_i = x] = \beta' x_i + \tau_0$$

then, you will have a model

$$y_i = \beta' x_i + \tau D_i + \epsilon_i$$

- You can think the matching estimator as controlling for omitted variable bias by adding (many) covariates (control variables) x_i.

Approach 2: $M-$Nearest Neighborhood Matching

- Idea: Find the counterpart in other group that is close to me.
- Define $\hat{y}_i(0)$ and $\hat{y}_i(1)$ be the estimator for (hypothetical) outcomes when treated and not treated.

$$\hat{y}_i(0) = \begin{cases} y_i & if\ D_i = 0 \\ \frac{1}{M}\sum_{j \in L_M(i)} y_j & if\ D_i = 1 \end{cases}$$

- $L_M(i)$ is the set of M individuals in the opposite group who are "close" to individual i
- Several ways to define the distance between X_i and X_j, such as

$$dist(X_i, X_j) = ||X_i - X_j||^2$$

- Need to choose (1) M and (2) the measure of distance

Approach 3: Propensity Score Matching

- Use propensity score $P(D_i = 1|X_i = x)$ as a distance to define who is the closest to me.
- Step 1: Estimate propensity score function by logit or probit using a flexible function of X_i.
- Step 2: Calculate the propensity score for each observation. Use it to define the pair.