Advanced Econometrics: OLS

Huan Deng

April 29, 2025

Outline

Overview

Next, we will continue our discussion of RCT focusing on estimation and inference

1. With random assignment, we can identify the average treatment effects under the assumption of SUTVA
2. We will consider probably the most important estimation method: OLS
3. We also want to quantify the uncertainty of our estimates, which require statistical inference
4. We will consider two types of inference methods: population-based inference (sampling-based uncertainty), and Fisherian approach (designed-based uncertainty)
5. Today we will discuss estimation, inference will be the next topic.

Linear Regressions

Linear Regressions

CEF and BLP

Conditional Expectation

- The conditional expectation of a random variable Y given another random variable X is:

$$E[Y|X]$$

- It represents the expected value of Y given the information contained in X.
- Defined as:

$$E[Y|X] = \int_{-\infty}^{\infty} yf(y|X)dy$$

where $f(y|X)$ is the conditional density of Y given X.

Law of Iterated Expectations (Thm 2.2 from Hansen)

Theorem

If $E|Y| < \infty$, then for any random vectors X_1 and X_2,

$$E[E[Y|X_1, X_2]|X_1] = E[Y|X_1]$$

- "The smaller information set wins"
- A special case:

$$E[E[Y|X]] = E[Y]$$

Conditioning Theorem (Thm 2.3 from Hansen)

Theorem

If $E|Y| < \infty$*, then:*

$$E[g(X)Y|X] = g(X)E[Y|X]$$

If in addition $E|g(X)| < \infty$*, then:*

$$E[g(X)Y] = E[g(X)E[Y|X]]$$

- These results show how expectations behave when multiplied by functions of the conditioning variable.
- They are useful for deriving properties of regression functions and expectations in econometrics.

CEF Error Definition

- The Conditional Expectation Function (CEF) is defined as:

$$m(X) = E[Y|X]$$

- The CEF error is the deviation from the expected value:

$$e = Y - E[Y|X]$$

- By definition, the error has zero mean conditional on X (Mean Independence):

$$E[e|X] = 0$$

Properties of CEF Error

- The CEF error is uncorrelated with any function of X:

  $$E[g(X)e] = 0$$

  for any measurable function $g(X)$.
- Note that Mean Independence doesn't imply that e and X are independent
- "X and e are independently distributed" is much stronger than Mean Independence

Conditional Expectation as Best Predictor

- Conditional expectation minimizes mean squared error:

$$E[(Y - m(X))^2] \leq E[(Y - g(X))^2]$$

for any other function $g(X)$.

- Proof:

$$\begin{aligned}
E\left[(Y - g(X))^2\right] &= E\left[(e + m(X) - g(X))^2\right] \\
&= E[e^2] + 2E[e(m(X) - g(X))] + E\left[(m(X) - g(X))^2\right] \\
&= E[e^2] + E\left[(m(X) - g(X))^2\right] \\
&\geq E[e^2] \\
&= E\left[(Y - m(X))^2\right].
\end{aligned}$$

Definition of Conditional Variance

- The conditional variance of Y given X is defined as:

$$\sigma^2(x) = \text{Var}(Y|X = x) = E[(Y - E[Y|X])^2|X = x]$$

- When $\sigma^2(x) = \sigma^2$, the error is homoskedastic; heteroskedastic, otherwise.
- The variance of Y can be decomposed as:

$$\text{Var}(Y) = E[\text{Var}(Y|X)] + \text{Var}(E[Y|X])$$

- The two terms on the right hand side are called "within group variance", and "between group variance".

Definition of Best Linear Predictor

- The CEF function is the best predictor for Y, but we don't know the functional form of $m(x)$. Let's take a step back and ask what is the Best Linear Predictor.
- The best linear predictor of Y given X is a linear function $X'\beta$ that minimizes the mean squared error:

$$\min_{\beta} E[(Y - X'\beta)^2]$$

- The optimal coefficient β is given by:

$$\beta = (E[XX'])^{-1}E[XY]$$

Properties of Best Linear Predictor

- The best linear predictor satisfies the orthogonality condition:

$$E[X(Y - X'\beta)] = 0$$

- This means the residual $e = Y - X'\beta$ is uncorrelated with X.
- The best linear predictor is the closest linear approximation to $E[Y|X]$ in terms of mean squared error.

Linear Regressions

The Algebra of OLS

Samples

- We have introduced the BLP coefficient:

$$\beta = (E[XX'])^{-1}E[XY]$$

- Now we want to estimate the above coefficient using a sample $\{(Y_i, X_i) : i = 1, ..., N\}$:

$$\hat{\beta} = (\sum_{i=1}^{N} X_i X_i')^{-1} \sum_{i=1}^{N} X_i Y_i$$

- Assumption: IID samples from population F.
- IID: independent and identical distribution

Notation

- Now we discuss how we get $\hat{\beta}$
- Consider an IID sample (Y_i, X_i).
- Y_i is a scalar, the outcome/dependent variable.
- X_i is a $(K+1) \times 1$ vector of independent variables:

$$X_i = [1, X_{i1}, X_{i2}, \ldots, X_{iK}]'$$

- Covariates are indexed by X_{ij} where i indexes observation and j indexes variables.

A Linear Model of Y and X

- We express Y_i as a linear function of X_i:

$$Y_i = X_i'\beta + \epsilon_i$$

- Write it out:

$$Y_i = \beta_0 + X_{i1}\beta_1 + X_{i2}\beta_2 + \cdots + X_{iK}\beta_K + \epsilon_i$$

- Here, β is a $(K+1) \times 1$ vector of unknown population parameters.
- β_0 is the called the intercept, and $\beta_1, \ldots, \beta_K$ are called slope.

Matrix Notation

- We can stack equations for all observations and write in matrix form:
$$Y = X\beta + \epsilon$$
- When working with matrix, pay attention to dimensions:
  - Y is $N \times 1$
  - X is $N \times (K+1)$
  - β is $(K+1) \times 1$
  - ϵ is $N \times 1$

OLS Estimator

- The least squares estimator is the solution to the below mean squared error problem:

$$\hat{\beta} = \arg\min_{\beta} \frac{1}{N} \sum_{i=1}^{n} (Y_i - X_i'\beta)^2$$

- We can express the OLS estimator in scalar form:

$$\hat{\beta} = \left(\frac{1}{N} \sum_{i} X_i X_i' \right)^{-1} \left(\frac{1}{N} \sum_{i} X_i Y_i \right)$$

- In vector/matrix form:

$$\hat{\beta} = (X'X)^{-1} X'Y$$

Least Squares Estimator

- $\hat{\beta}$ is a $(K+1) \times 1$ vector of estimats.
- Sometimes denoted as $\hat{\beta}_{OLS}$ when compared with other estimates.
- It is a statistic (function of data) and a random variable with a distribution.
- The Least Squares Estimator is the most popular estimator in Applied Econometrics

Orthogonality in OLS

- Define $\hat{e} = Y - X\hat{\beta} = Y - \hat{Y}$, which is called the "residual"
- **Theorem:** X is orthogonal to the residual $\hat{e}$ in OLS regression.

$$X'\hat{e} = 0_{K+1}$$

- Every covariate in X is orthogonal to the residuals.
- This means that OLS residuals have zero correlation with the regressors.

Proof of Orthogonality

Proof:

$$\begin{aligned}
X'\hat{e} &= X'(Y - X\hat{\beta}) \\
&= X'(Y - X(X'X)^{-1}X'Y) \\
&= X'Y - X'X(X'X)^{-1}X'Y \\
&= (X' - X'X(X'X)^{-1}X')Y \\
&= 0_{K+1}
\end{aligned}$$

OLS Decomposition

Definition: The total variation in Y can be decomposed as:

$$\text{(SST)} = \text{(SSE)} + \text{(SSR)}$$

$$Y'Y = \hat{Y}'\hat{Y} + \hat{e}'\hat{e}$$

- $Y'Y$ (SST): total variation in Y.
- $\hat{Y}'\hat{Y}$ (SSE): variation explained by the regression.
- $\hat{e}'\hat{e}$ (SSR): unexplained variation (Residual Sum of Squares).

R-Squared: Definition

Definition:

$$R^2 = 1 - \frac{\sum \hat{e}_i^2}{\sum (Y_i - \bar{Y})^2}$$

- $R^2 \in [0, 1]$, measuring the proportion of variance in Y explained by X.
- $R^2 = 1$ if all residuals are zero (perfect fit).
- $R^2 = 0$ if all the slopes are zero.

Interpreting R^2

Properties:

- R^2 is a crude measure of regression fit.
- It always increases when more regressors are added.
- Adjusted R^2 accounts for added variables and can decrease if they don't improve fit.

Projection Matrix

- The projection matrix:

$$P = X(X'X)^{-1}X'$$

- Projects Y onto the space spanned by X:

$$PY = X\hat{\beta} = \hat{Y}$$

- P has some nice properties: for example, $PX = X$, symmetric $P = P'$, idempotent $PP = P$

Annihilator Matrix

- The annihilator matrix:

$$M = I - P$$

- It generates residuals:

$$MY = \hat{e}$$

- It also have some nice properties: $MX = 0$, symmetric $M = M'$, idempotent $MM = M$

Regression Components

- Consider partitioning the regressor matrix $X = [X_1, X_2]$ and the coefficient vector $\beta = (\beta_1, \beta_2)$.
- The regression model can be written as:

$$Y = X_1\beta_1 + X_2\beta_2 + e$$

- The least squares estimator is obtained by minimizing the sum of squared errors:

$$(\beta_1, \beta_2) = \arg\min_{\beta_1,\beta_2} SSR(\beta_1, \beta_2)$$

- where the SSR is given by:

$$\text{SSR}(\beta_1, \beta_2) = (Y - X_1\beta_1 - X_2\beta_2)'(Y - X_1\beta_1 - X_2\beta_2)$$

Estimating β_1

- The estimator β_1 is found by concentrating out β_2:

$$\hat{\beta}_1 = \arg\min_{\beta_1} \left(\min_{\beta_2} \mathsf{SSR}(\beta_1, \beta_2) \right)$$

- This represents a nested minimization problem:
  1. First, minimize the inner SSR over β_2 at any given β_1.
  2. Then, find the optimal β_1 that minimizes the resulting function.

Estimating β_1

- The inner problem is just a regression of $Y - X_1\beta_1$ on X_2.
- So the solution for the inner problem is:

$$\arg\min_{\beta_2} \mathrm{SSR}(\beta_1, \beta_2) = (X_2'X_2)^{-1}X_2'(Y - X_1\beta_1)$$

- **Rewriting in terms of residuals:**

$$Y - X_1\beta_1 - X_2(X_2'X_2)^{-1}X_2'(Y - X_1\beta_1)$$

$$= (M_2Y - M_2X_1\beta_1)$$

$$= M_2(Y - X_1\beta_1)$$

where

$$M_2 = I_N - X_2(X_2'X_2)^{-1}X_2'$$

Estimating β_1

- The inner problem has a minimum:

$$\min_{\beta_2} \mathrm{SSR}(\beta_1, \beta_2) = (Y - X_1\beta_1)' M_2 M_2 (Y - X_1\beta_1)$$

$$= (Y - X_1\beta_1)' M_2 (Y - X_1\beta_1)$$

  Since M_2 is idempotent, the second equality holds.
- Substituting back, we obtain:

$$\hat{\beta}_1 = \arg\min_{\beta_1} (Y - X_1\beta_1)' M_2 (Y - X_1\beta_1)$$

  which gives the solution:

$$\hat{\beta}_1 = (X_1' M_2 X_1)^{-1} (X_1' M_2 Y)$$

Estimating β_2

- Similarly, the estimator for β_2 is given by:

$$\beta_2 = (X_2' M_1 X_2)^{-1} X_2' M_1 Y$$

where M_1 is the annihilator matrix:

$$M_1 = I_n - X_1 (X_1' X_1)^{-1} X_1'$$

- $M_1 Y$ represents the residual from regressing Y on X_1.
- $M_1 X_2$ represents the residual from regressing X_2 on X_1.

Theorem: Expression for Regression Components

Theorem: The least squares estimator for β_1 and β_2 has the algebraic solution:

$$\beta_1 = (X_1' M_2 X_1)^{-1} X_1' M_2 Y$$

$$\beta_2 = (X_2' M_1 X_2)^{-1} X_2' M_1 Y$$

where M_1 and M_2 are the corresponding annihilator matrices.

Interpretation

- The estimates β_2 can be computed through a two-step process:
  1. Regress X_2 on X_1 and obtain residuals, and regress Y on X_1 and obtain residuals.
  2. Perform OLS on the residuals.
- This is called Frisch-Waugh-Lovell (FWL) Theorem
- This is especially useful if we mainly care about some "causal parameter" and other parameters are just nuisance parameters (can be high-dimensional)

Application: Regression for RCT

Estimating ATE

- Recall that
  $ATE = E[\tau_i] = E[Y_i(1) - Y_i(0)] = E[Y_i(1)] - E[Y_i(0)]$
- How can we estimate the ATE with a sample of (Y_i, D_i)?
- Moment Estimator is a natural choice

Moment Estimators

- Population expectation and variance of a random variable:

$$\mu = E[Y], \sigma^2 = E[Y^2] - E[Y]^2$$

- Sample mean estimator:

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} Y_i$$

- Variance estimation:

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} Y_i^2 - \left(\frac{1}{n} \sum_{i=1}^{n} Y_i \right)^2$$

Difference-in-Means Estimator

- Denote $N_1 = \Sigma D_i$ and $N_0 = \Sigma(1 - D_i)$ as the number of units in treatment and control groups
- Then we define the DM estimator as:

$$\hat{\tau}_{DM} = \frac{1}{N_1} \sum_{D_i=1} Y_i - \frac{1}{N_0} \sum_{D_i=0} Y_i$$

- We use the sample mean for the treatment (control) group as an estimate for $E[Y_i(1)]$ ($E[Y_i(0)]$)

Difference-in-Means Estimator

- The DM estimator is unbiased. For $w \in \{0, 1\}$:

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{N_w}\sum_{D_i=w} Y_i\right] &= \mathbb{E}\left[Y_i \mid D_i = w\right] (\text{IID}) \\
&= \mathbb{E}\left[Y_i(w) \mid W_i = w\right] (SUTVA) \\
&= \mathbb{E}\left[Y_i(w)\right] \quad (\text{random assignment})
\end{aligned}$$

Regression for RCT

- We can also get the above difference-in-means estimates by running a regression:

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

- We can derive the $\hat{\beta}_1$ as:

$$\begin{aligned}
\hat{\beta}_1 &= \frac{\sum_{i=1}^{N}(D_i - \bar{D})(Y_i - \bar{Y})}{\sum_{i=1}^{N}(D_i - \bar{D})^2} \\
&= \frac{\sum_{D_i=1}(1 - \frac{N_1}{N})(Y_i - \bar{Y}) + \sum_{D_i=0}(0 - \frac{N_1}{N})(Y_i - \bar{Y})}{N_1 N_0 / N} \\
&= \hat{\tau}_{DM}
\end{aligned}$$