Workshop on AI Meets Econometrics

From Machine Learning to Deep Learning

Fei Wang

Management and Microeconomics
Goethe Universität Frankfurt

25 May 2022

Roadmap

INTRODUCTION

The Big Picture

How will we demystify machine learning and deep learning?

1. Review of BLUE
2. Review of Causal Inference
3. From Econometrics to Machine Learning
4. From Machine Learning to Deep Learning
5. A Concrete Example: NLP

https://en.wikipedia.org/wiki/Deep_learning

REVIEW OF BLUE

Review of BLUE

For the following multivariate linear regression model,

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_t \\ \vdots \\ y_N \end{bmatrix} = \begin{bmatrix} 1 & x_{12} & \cdots & x_{1k} \\ 1 & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \\ 1 & x_{t2} & \cdots & x_{tk} \\ \vdots & \vdots & \vdots & \\ 1 & x_{N2} & \cdots & x_{Nk} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_t \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_t \\ \vdots \\ u_N \end{bmatrix}$$

or simply:

$$Y = X\beta + u \tag{1}$$

where, Y is a $N \times 1$ vector, X is a $N \times k$ matrix, and β is a $k \times 1$ vector. With assumptions:

$$E(u) = 0 \qquad V(u) = \sigma^2 I$$

X is fixed   X has full rank k (or $k < N$)

Review of BLUE: Assumptions Matter

The model is built up for **population**:

$$Y = X\beta + u$$

For each **sample**, we use OLS method to estimate $\hat{\beta}_1, \cdots, \hat{\beta}_k$ that minimize the sum of **squared residuals**:

$$S = \sum_{i=1}^{N} \hat{u_t}^2 = (Y - X\hat{\beta})'(Y - X\hat{\beta})$$

and yields the following estimator in matrix form:

$$\hat{\beta} = (X'X)^{-1}X'Y \tag{2}$$

Review of BLUE: Assumptions Matter (1)

OLS estimator $\hat{\beta}$ is the Best **Linear Unbiased** Estimator because we assume $E(u) = 0$. With the following population model and estimator:

$$
\begin{aligned}
Y &= X\beta + u \\
\hat{\beta} &= (X'X)^{-1}X'Y = (X'X)^{-1}X'(X\beta + u) \\
\hat{\beta} &= \beta + u \\
\rightarrow E(\hat{\beta}) &= E(\beta + u) = E(\beta) + E(u) \\
E(\hat{\beta}) &= \beta
\end{aligned}
$$

The Gauss-Markov theorem also states that the variance of OLS estimator $\hat{\beta}$ has the **smallest variance**:

$$V(\hat{\beta}) = \sigma^2(X'X)^{-1},$$

which make it become the **Best Linear Unbiased Estimator** (BLUE).

Review of BLUE: Assumptions Matter (2)

We reviewed one assumption The variance of $\hat{\beta}$ is:

$$V(\hat{\beta}) = \sigma^2 (X'X)^{-1},$$

However, σ^2 is unknown as it is assumed for the **population** model. We have to **estimate** σ^2. By imposing the **second assumption**:

$$V(u) = \sigma^2 I$$

we could have the following equation:

$$E\left(\frac{\hat{u}'\hat{u}}{N-k}\right) = \sigma^2$$

Therefore, an **unbiased** estimator of σ^2 is provided by

$$\hat{\sigma}^2 = \frac{\hat{u}'\hat{u}}{N-k}$$

Review of BLUE: Assumptions Matter (3)

Now, we check the third assumption: X is fixed (or non-stochastic), which could be interpreted as 'distribution is fixed' rather than stochastic (for example, a random walk). With this assumption, we could safely add one more assumption:

$$u \sim N(0, \sigma^2)$$

The above assumption **was not** included in the model that assumes:

1. $E(u) = 0$
2. $V(u) = \sigma^2 I$
3. X is fixed
4. X has full rank (k ¡ N)

When X is fixed and error term u (for population model) follows the normal distribution

$$u \sim N(0, \sigma^2)$$

, we could have:

$$\hat{\beta} \sim N(\beta, \sigma^2 (X'X)^{-1})$$

Review of BLUE: Assumptions Matter (3)

One more step! Now, we have the following distribution:

$$\hat{\beta} \sim N(\beta, \sigma^2 (X'X)^{-1})$$

However, we don't know σ^2. After replacing σ^2 with the unbiased estimator $\hat{\sigma}^2$, it turned out that the hypothesis such as $\beta_i = \beta_{i0}$ could be tested by comparing the test statistic:

$$\frac{\hat{\beta}_i - \beta_{i0}}{\sqrt{\hat{\sigma}^2 \alpha_{ii}}}$$

with **critical values** from the t_{N-k} distribution, where α_{ii} is the ii's element of $(X'X)^{-1}$.

Review of BLUE: Assumptions Matter (4)

The final assumption: X is full rank ($k < N$) Try to solve the following different cases:

- $\beta_1 + 2\beta_2 = 3$, infinite solutions (k = 2, N = 1)
- unique solutions ($k = 2, N = 2$, full rank)

$$\beta_1 + 2\beta_2 = 3$$
$$\beta_1 - \beta_2 = 5$$

- unique solutions ($k < N$) too

$$\underset{N \times 1}{\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}} = \underset{N \times k}{\begin{bmatrix} 1 & x_{12} & \cdots & x_{1k} \\ 1 & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \\ 1 & x_{N2} & \cdots & x_{Nk} \end{bmatrix}} \underset{k \times 1}{\begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix}} + \underset{N \times 1}{\begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_N \end{bmatrix}}$$

Review of BLUE: Summary

- $E(u) = 0$ gives us Linear Unbiased Estimator
- $V(u) = \sigma^2 I$ gives us BLUE
- $V(u) = \sigma^2 I$ gives us $\hat{\sigma}^2$ too
- X is fixed and $u \sim N(0, \sigma^2)$ enable us to do t-test
- X is full rank makes sure that we could solve equations

Why bother to do all those things?

- collecting data is not easy (sample size is often small)
- making inference on population needs the assumption of distribution
- having a framework of conducting hypothesis test
- confidence in assumed distribution gives the confidence in inference
  - it is often the case
  - Law of Large Numbers (LLN) and Central Limit Theorem (CLT) guarantees this