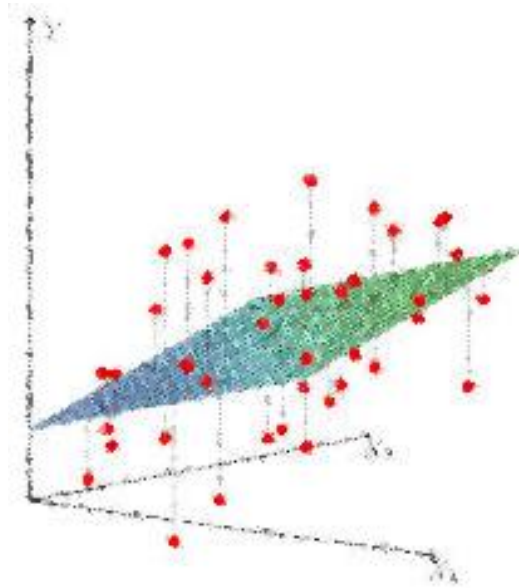




# Module 3 – Linear Regression Models

## Econometrics Approach

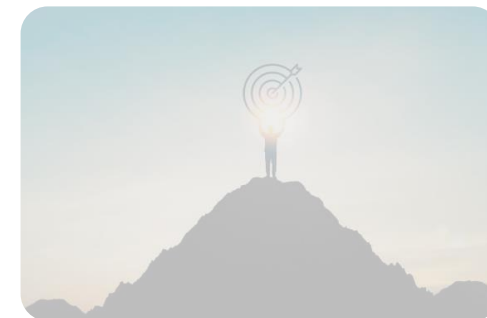
---





# Class Modules

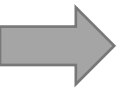
- Module 1- Introduction to Deep Learning
- Module 2- Setting up Machine Learning Environment
- **Module 3- Linear Regression (Econometrics approach)**
- Module 4- Machine Learning Fundamentals
- Module 5- Linear Regression (Machine Learning approach)
- Module 6- Penalized Regression (Ridge, LASSO, Elastic Net)
- Module 7- Logistic Regression
- Module 8- K-Nearest Neighbors (KNN)
- Module 9- Classification and Regression Trees (CART)
- Module 10- Bagging and Boosting
- Module 11- Dimensionality Reduction (PCA)
- Module 12- Clustering (KMeans – Hierarchical)



# ➔ What is Econometrics?

---

- Econometrics is the branch of economics that develops and uses **statistical methods** for **estimating economic relationships**
- Typical goals of econometrics analysis are:
  - **Estimating** relationships between random variables
  - **Testing** hypothesis
  - **Predicting** / Forecasting random variables



# Steps in Econometrics analysis

---

- 1) Specifying the **regression** model
- 2) Collecting data
- 3) Quantify the model

Example:



# The Multiple Regression Model (MRM)

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \dots + \beta_k x_k + u$$

- In econometrics language,  $\beta_s$  are the coefficients and  $u$  is the error term.

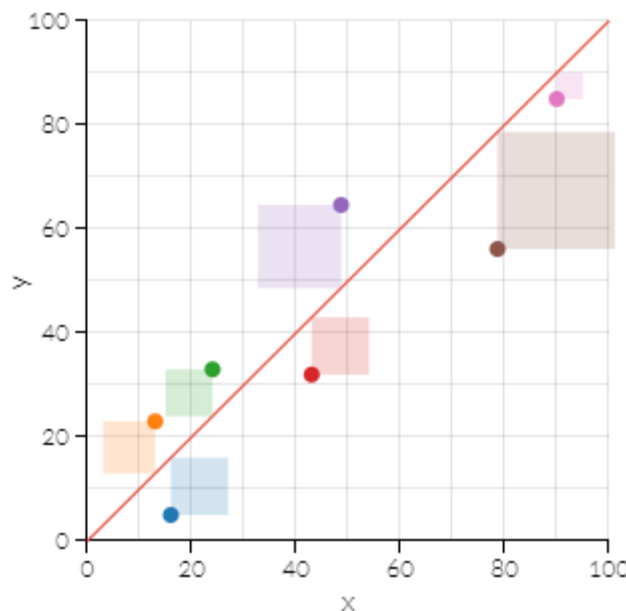
Y	X
Dependent variable	Independent variable
Explained variable	Explanatory variable
Response variable	Control variable
Predicted variable	Predictor variable
Regressand	Regressor

- How to estimate the coefficients? It's all about the **error term u**

# ➔ Estimating the model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \dots + \beta_k x_k + u$$

- OLS (Ordinary Least Squared) is **one way** to estimate the coefficients.
- In order to **estimate** this model (finding out the  $\beta_s$ ) , we need to make some **assumptions**.



# → Interpreting the model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \dots + \beta_k x_k + u$$

- Interpreting  $\beta_j$ 

if  $x_j$  increases by 1 unit, holding everything else constant, on average,  $y$  will increase by  $\beta_j$  units.
- In order to **interpret** this model (interpreting the  $\beta_s$ ) , we need to make some **assumptions**.

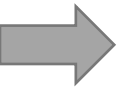
# → Correlation vs Causation

---

- Correlation refers to the **linear** relationship between two variables, and how they change together.
- Causation refers to the **cause and effect**, where the one event is a result of another event .
- Regression analysis cannot prove causality!
- Given some **assumptions**, we **hope** to get causality with statistical significance.
- What are these assumptions?

## The Gauss-Markov assumptions

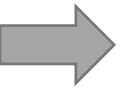




# The Gauss-Markov assumptions (cont'd)

---

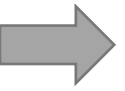
- **Assumption 1:** Linearity in parameters
- **Assumption 2:** Random Sampling
- Examples:



# The Gauss-Markov assumptions (cont'd)

---

- **Assumption 3:** No perfect collinearity and  $\text{var}(x) \neq 0$
- Examples:



# The Gauss-Markov assumptions (cont'd)

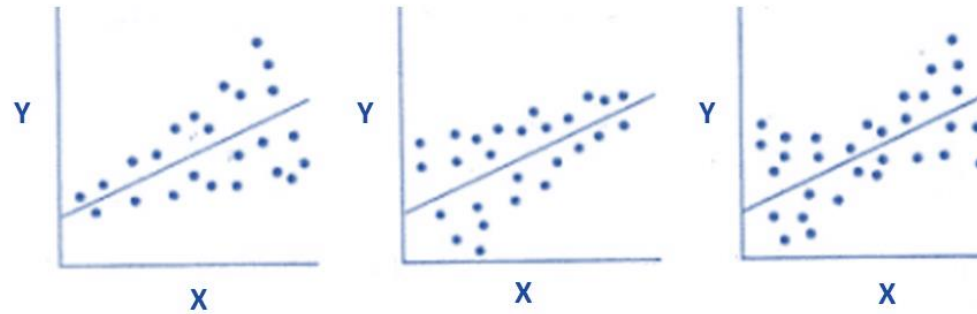
---

- **Assumption 4:** Zero Conditional Mean. Given any values of  $X$ , the errors are on average zero (conditional expectation).  $E(u|X) \neq 0$
- **Endogeneity** violates this assumption!  $Corr(X, u) \neq 0$
- Examples:



# The Gauss-Markov assumptions (cont'd)

- **Assumption 5:** Homoskedasticity (same conditional variance) :  $\text{var}(u|X) \neq \sigma^2$
- Examples:





# The Gauss-Markov assumptions (summary)

OLS estimators are unbiased

- 1) Assumption 1: Linearity in parameters
- 2) Assumption 2: Random Sampling
- 3) Assumption 3: No perfect collinearity and  $\text{var}(x) \neq 0$
- 4) Assumption 4: Zero Conditional Mean
- 5) Assumption 5: Homoskedasticity

There is a formula for variance of OLS estimators

# → The scope of our course!

---

- If any of the Gauss-Markov assumptions are not met, it is important to exercise caution when interpreting the results of the econometric model, as the model's **predictions** and **parameter estimates** may be unreliable.
- Statistical tests and tools can be used to verify any of these assumptions, but it's beyond the scope of this course.

# Statistical Inference

## Hypothesis Testing

(quick review)

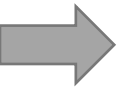


# Statistical inference in the regression model

---

- So far, given the GMA, we know something about the **expected value** and the **variance** of OLS estimators. What about its **distribution**?
- We need one more **assumption! Oh no!**
- **Assumption 6:** the error terms are normally distributed,  $u \sim N(0, \sigma^2)$
- Assumptions 1 through 6 is called, Classical Linear Model (CLM) assumptions.
- Good news: if the sample size is large **enough**, we can relax the normality assumption (because of central limit theorem)





# Theorem: t distribution for the estimators

---

- Under the CLM assumptions,

$$\frac{\hat{\beta}_j - \beta_j}{se(\hat{\beta}_j)} \sim t_{n-k-1}$$

- Now we can do hypothesis testing! Yay!
- Review hypothesis testing if needed.



# Evaluation Metrics

- Now let's focus on the **performance** aspect of linear regression models:

$$R^2 = 1 - \frac{\sum (y_i - \hat{y})^2}{\sum (y_i - \bar{y})^2} = 1 - \frac{SS_{residuals}}{SS_{total}}$$

$$Adjusted R^2 = 1 - (1 - R^2) * \frac{n - 1}{n - k - 1}$$

$$MAE = \frac{1}{N} \sum_{i=1}^N |y_i - \hat{y}|$$

$$MAPE = \frac{100\%}{n} \sum \left| \frac{y - \hat{y}}{y} \right|$$

$$MSE = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y})^2$$

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{N} \sum_{i=1}^N (y_i - \hat{y})^2}$$



# Class Modules

- ✓ Module 1- Introduction to Deep Learning
- ✓ Module 2- Setting up Machine Learning Environment
- ✓ Module 3- Linear Regression (Econometrics approach)
- Module 4- Machine Learning Fundamentals
- Module 5- Linear Regression (Machine Learning approach)
- Module 6- Penalized Regression (Ridge, LASSO, Elastic Net)
- Module 7- Logistic Regression
- Module 8- K-Nearest Neighbors (KNN)
- Module 9- Classification and Regression Trees (CART)
- Module 10- Bagging and Boosting
- Module 11- Dimensionality Reduction (PCA)
- Module 12- Clustering (KMeans – Hierarchical)

