Introductory Econometrics

Simple Linear Regression Model (I)

Yaohan Chen

School of Big Data and Statistics, Anhui University

Spring, 2025

Relationship

- Relationship between variables
  - Deterministic

$$\text{Area of a Circle} = f(\pi, \text{radius}) = \pi \cdot \text{radius}^2$$

  - Correlation

$$\begin{aligned} \text{yield} &= f(\text{temperature}, \text{percipitation}, \text{sunshine}, \text{fertilizer}) \\ &\quad + \text{randomness} \end{aligned}$$

Correlation Analysis

- Linear correlation and Non-linear correlation
- Linear population correlation

$$\rho_{XY} = \frac{\operatorname{Cov}(X, Y)}{\sqrt{\operatorname{Var}(X)\operatorname{Var}(Y)}}$$

- Linear sample correlation

$$r_{XY} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

Regression Analysis

- explained variable (or dependent variable), Y.
- explanatory variable (or independent variable), X.
- Regression: to recover relationship.
  - to explain Y in terms of X.
  - to study how Y varies with changes in X.
  - to predict Y for given values of X.

  Example: By how changes the hourly wage for additional year of schooling ?
- Regression analysis lays the methodological foundation for Econometrics.

Population Regression Model

Example:

- Community with 99 families.
    - Y: monthly expenditure.
    - X: monthly income.
- Dissecting 99 families into 10 groups.
- Can we predict expenditure if we know income ?

表 2.1.1　某社区家庭每月收入与消费支出统计表

	每月家庭可支配收入X（元）									
	800	1100	1400	1700	2000	2300	2600	2900	3200	3500
每月家庭消费支出Y（元）	561	638	869	1023	1254	1408	1650	1969	2090	2299
	594	748	913	1100	1309	1452	1738	1991	2134	2321
	627	814	924	1144	1364	1551	1749	2046	2178	2530
	638	847	979	1155	1397	1595	1804	2068	2266	2629
		935	1012	1210	1408	1650	1848	2101	2354	2860
		968	1045	1243	1474	1672	1881	2189	2486	2871
			1078	1254	1496	1683	1925	2233	2552	
			1122	1298	1496	1716	1969	2244	2585	
			1155	1331	1562	1749	2013	2299	2640	
			1188	1364	1573	1771	2035	2310		
			1210	1408	1606	1804	2101			
				1430	1650	1870	2112			
				1485	1716	1947	2200			
						2002				
共计	2420	4950	11495	16445	19305	23870	25025	21450	21285	15510

Population Regression Model

Analysis:

- For specific income level, i.e. given X, expenditure level may be different, i.e. Y varies. Why ?

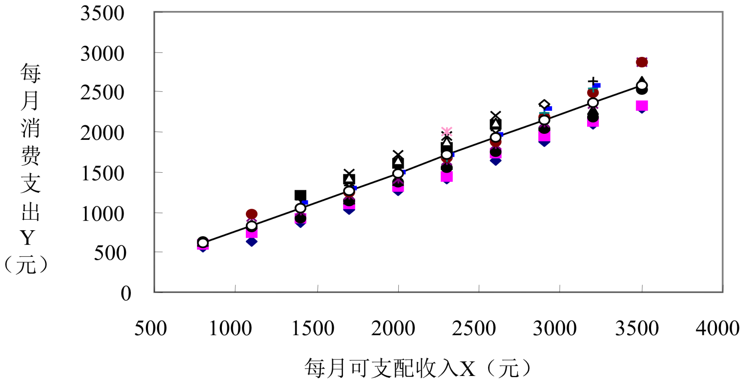
Population Regression Model

Analysis:

- For specific income level, i.e. given X, expenditure level may be different, i.e. Y varies. Why ?
- We can depict the uncertainty using **conditional distribution**. For instance,

$$\mathrm{P}(Y = 561 \mid X = 800) = 1/4$$

Income X	800	1100	1400	1700	2000	2300	2600	2900	3200	3500
Conditional Probability	1/4	1/6	1/11	1/13	1/13	1/14	1/13	1/10	1/9	1/6
Conditional Mean $\mathrm{E}(Y \mid X)$	605	825	1045	1265	1485	1705	1925	2145	2365	2585



- For given $X = X_i$, we calculate the **conditional mean** of Y.

$$\mathrm{E}(Y \mid X = X_i)$$

Population Regression Model

- **Population Regression Line** is a line depicting the conditional expectation of explained variable Y conditional on X. It is referred to as **Population Regression Curve**.
- The function associated with the population regression line is called the **Population Regression Function, PRF**

$$\mathrm{E}\left(Y \mid X\right) = f\left(X\right).$$

- What is the functional form of $f(X)$?
- If $f(X)$ is linear,

$$\mathrm{E}\left(Y \mid X\right) = \beta_0 + \beta_1 X.$$

where β_0 and β_1 are called the **regression coefficients**.

Population Regression Model

- PRF describes how Y varies across X **on average**. But the Y we observe are **random variables**.
- How to model the randomness ?

$$u = Y - \mathrm{E}(Y \mid X).$$

where u depicts the **deviation** of Y relative to $\mathrm{E}(Y \mid X)$.

- u is **unobserved** random variable, referred to as the **stochastic error** or **stochastic disturbance**.
- In general, we have **Population Regression Model**

$$Y = \mathrm{E}(Y \mid X) + u.$$

Population Regression Model

- Bivariate linear regression model

$$Y = \beta_0 + \beta_1 X + u.$$

  - systematic part: $\beta_0 + \beta_1 X$, or deterministic part.
  - nonsystematic part: u.

Why do we need to include u ?

Population Regression Model

- Bivariate linear regression model

$$Y = \beta_0 + \beta_1 X + u.$$

  - systematic part: $\beta_0 + \beta_1 X$, or deterministic part.
  - nonsystematic part: u.

Why do we need to include u ?

- The unknown potential determinants.
- Missing data for various reasons.
- Insignificant determinants.

Population Regression Model

- Measurement errors.
- Model misspecification errors.
  - stochastic error u can only partially capture errors of this kind.
- Other randomness.

Sample Regression Function

- We observe the **sample** of random variable, say $\{Y_i\}_{i=1}^{n}$ of Y. How to use the sample information to approximate the population information ?

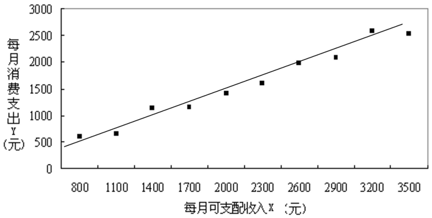
X	800	1100	1400	1700	2000	2300	2600	2900	3200	3500
Y	638	935	1155	1254	1408	1650	1925	2068	2266	2530

Sample Regression Function

- We observe the **sample** of random variable, say $\{Y_i\}_{i=1}^n$ of Y. How to use the sample information to approximate the population information ?

X	800	1100	1400	1700	2000	2300	2600	2900	3200	3500
Y	638	935	1155	1254	1408	1650	1925	2068	2266	2530

- **Scatter Diagram**:



Sample Regression Function

- A line that fits the scatters is **Sample Regression Line**.
- **Sample Regression Function, SRF** is the functional form associated with the sample regression line

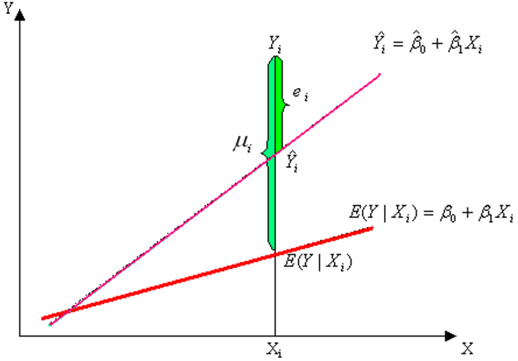
$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X.$$

- Sample Regression Function as the approximation of Population Regression Function.

$$\begin{aligned} Y &= \mathrm{E}(Y \mid X) + u \\ &= \beta_0 + \beta_1 X + u \\ &= \hat{\beta}_0 + \hat{\beta}_1 X + e \end{aligned}$$

Sample Regression Function

- e is referred to as **residual**.
- SRF and PRF



Regression: To Estimate PRF via SRF.

Recall: Bivariate Linear Regression Model

- Recall the bivariate linear regression model

$$Y = \beta_0 + \beta_1 X + u,$$

  where β_0 and β_1 are parameters to be estimated, and
  - β_0 is referred to as the **intercept**.
  - β_1 is referred to as the **slope**.
- We observe Y and X as **sample**,

$$\{(X_i, Y_i) : i = 1, 2, \ldots, n\}$$

- **For each i**, bivariate linear regression model suggests

$$Y_i = \beta_0 + \beta_1 X_i + u_i.$$

Ordinary Least Square Estimation

- Estimating Sample Regression Function is equivalent to obtain estimation of β_0 and β_1. Alternatively, how to establish the connection between sample and β_0 and β_1.

Ordinary Least Square Estimation

- Estimating Sample Regression Function is equivalent to obtain estimation of β_0 and β_1. Alternatively, how to establish the connection between sample and β_0 and β_1.
- Among all the available estimation methods, we first consider obtaining $\hat{\beta}_0$ and $\hat{\beta}_1$ by minimizing quadratic loss function

$$Q = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2 = \sum_{i=1}^{n} \left[Y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 X_i\right)\right]^2$$

- Q measures the deviation as the sum of the squared deviations of Y_i to $\hat{Y}_i$.

Ordinary Least Square Estimation

- By taking derivatives with respect to $\hat{\beta}_0$ and $\hat{\beta}_1$ and the first order partial derivatives equal to 0

$$\begin{cases} \dfrac{\partial Q}{\hat{\beta}_0} = 0 \\ \dfrac{\partial Q}{\hat{\beta}_1} = 0 \end{cases}$$

- Solution

$$\begin{cases} \hat{\beta}_0 = \dfrac{\sum X_i^2 \sum Y_i - \sum X_i \sum X_i Y_i}{n \sum X_i^2 - \left(\sum X_i\right)^2} \\ \hat{\beta}_1 = \dfrac{n \sum Y_i X_i - \sum Y_i \sum X_i}{n \sum X_i^2 - \left(\sum X_i\right)^2} \end{cases}$$

Ordinary Least Square Estimation

- By Letting $x_i = X_i - \bar{X}$ and $y_i = Y_i - \bar{Y}$, we can rewrite $\hat{\beta}_0$ and $\hat{\beta}_1$ as

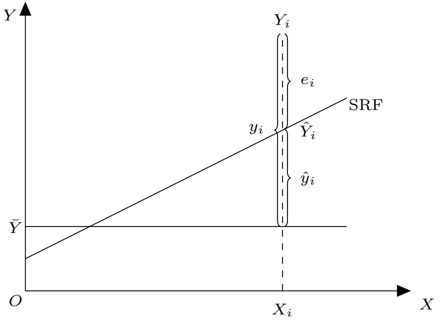
$$\begin{cases} \hat{\beta}_1 & = \dfrac{\sum x_i y_i}{\sum x_i^2} \\ \hat{\beta}_0 & = \bar{Y} - \hat{\beta}_1 \bar{X} \end{cases}$$

- $\hat{\beta}_0$ are $\hat{\beta}_1$ are called the **Ordinary Least Square estimator**, or OLS estimator.

Goodness-of-Fit

- How "good" SRF approximation is relative to PRF ?
- Decomposition of $y_i = Y_i - \bar{Y}$

$$y_i = Y_i - \bar{Y} = \underbrace{\left(Y_i - \hat{Y}_i\right)}_{e_i} + \underbrace{\left(\hat{Y}_i - \bar{Y}\right)}_{\hat{y}_i}$$



Goodness-of-Fit

- By taking the sum of squared y_i, we have

$$\sum y_i^2 = \sum \hat{y}_i^2 + \sum e_i^2 + 2 \sum \hat{y}_i e_i$$

and $\sum \hat{y}_i e_i = 0$ (why).

- TSS = ESS + RSS:

$$\underbrace{\sum y_i^2}_{\text{TSS}} = \underbrace{\sum \hat{y}_i^2}_{\text{ESS}} + \underbrace{\sum e_i^2}_{\text{RSS}}$$

where TSS refers to **Total Sum of Squares**, ESS refers to **Explained Sum of Squares**, and RSS refers to **Residual Sum of Squares**.

Goodness-of-Fit

- We can define the measure of goodness-of-fit as follows

$$R^2 = \frac{\text{ESS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}}.$$

  R^2 is called the **coefficient of determination**.

- We can calculate R^2 as follows due to the definition of ESS

$$R^2 = \hat{\beta}_1^2 \left(\frac{\sum x_i^2}{\sum y_i^2} \right).$$

  R^2 can be interpreted as the *fraction of sample variation of Y that is explained by X*.

Summary

- Regression.
- Population Regression Function and Sample Regression Function.
- Ordinary Least Square estimation and the corresponding derivation.
- Goodness of Fit.