# Econometrics

笔记使用中英双语。斜体为个人批注。翻译在括号中。

教材："Introduction to Econometrics (4th Edition)" by Stock, Watson

# Lecture 1

## 1.1 The population linear regression model (总体回归函数)

Linear regression lets us estimate the population regression line and its slope.

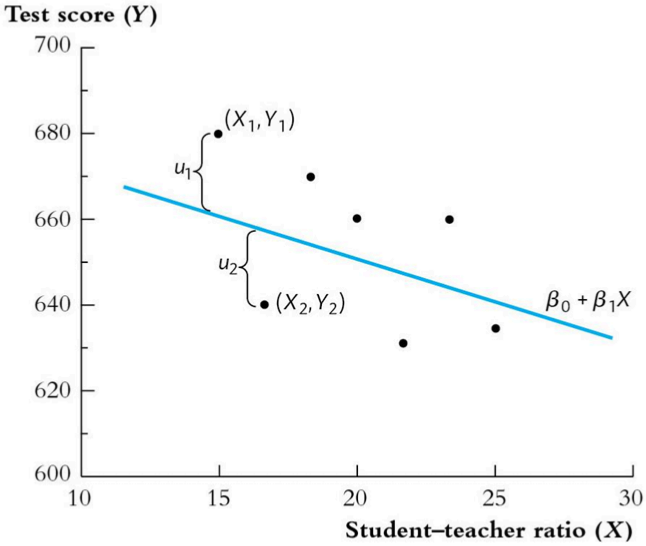
- The The population regression line is the **expected value** of $Y$ given $X$.
- The estimated regression can be used either for:
    - **causal inference** (learning about the causal effect on Y of a change in X)
    - **prediction** (predicting the value of Y given X, for an observation not in the data set)
- **Causal inference** and **prediction** place different requirements on the data – but both use the same regression toolkit.

Statistical, or econometric, inference about the slope entails

- Estimation:
    - How should we draw a line through the data to estimate the population slope?
        - Answer: ordinary least squares (OLS, 最小二乘法).
- Hypothesis testing
- Confidence intervals (置信区间)

$$Y_i = \beta_0 + \beta_1 X_i + u_i, i = 1, \ldots, n \tag{1}$$

- We have n observations, $(X_i, Y_i), i = 1, \ldots, n$.
- $X$ is the independent variable or regressor
- $Y$ is the dependent variable
- $\beta_0$ = intercept
- $\beta_1$ = slope
- $u_i$ = the regression error
- The regression error consists of omitted factors and error in the measurement of $Y$.




## 1.2 Derivation (推导) of OLS estimator (估计值) $\hat{\beta}_0$ and $\hat{\beta}_1$

Pick $\hat{\beta}_0$ and $\hat{\beta}_1$ to minimize the sum of the squared errors.

$$S = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)$$

We get

$$\begin{gathered} \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1\bar{X} \\ \hat{\beta}_1 = \frac{\sum_i^n(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_i^n(X_i - \bar{X})^2} \end{gathered} \tag{2}$$

The OLS predicted values $\hat{Y}_i$ and residuals $u_i$ are

$$\begin{gathered} \hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i \\ \hat{u}_i = Y_i - \hat{Y}_i \end{gathered} \tag{3}$$

## 1.3 Measures of Fit

Two regression statistics provide complementary measures of **how well the regression line "fits"** or explains the data.

### 1.3.1 The Regression $R^2$

It measures the fraction (比例) of the variance of $Y$ is explained by $X$. It ranges from 0 (no fit) to 1 (perfect fit).

$$R^2 = \frac{\text{ESS}}{\text{TSS}} = \frac{\sum_i^n(\hat{Y}_i - \bar{Y})^2}{\sum_i^n(Y_i - \bar{Y})^2} \tag{4}$$

- **TSS（Total Sum of Squares）**：$Y$的总变异（实际值与均值的偏离）。
- **ESS（Explained Sum of Squares）**：回归模型能解释的变异（预测值与均值的偏离）。
- **RSS（Residual Sum of Squares）**：$\sum_i^n \hat{u}_i{}^2$
  模型无法解释的残差异变（实际值与预测值的偏离）。

$$\text{TSS} = \text{ESS} + \text{RSS} \tag{5}$$

### 1.3.2 The Standard Error of the Regression (SER)

The SER measures the spread of the distribution of $u$. The SER is (almost) the sample standard deviation of the OLS residuals

$$
\begin{aligned}
\mathrm{SER} &= \sqrt{\frac{1}{n-2}\sum_i^n(\hat{u}_i-\bar{\hat{u}})^2} \\
&= \sqrt{\frac{1}{n-2}\sum_i^n \hat{u}_i{}^2}
\end{aligned} \tag{6}
$$

The second equality holds because $\bar{\hat{u}} = \frac{1}{n}\sum_i^n \hat{u}_i = 0$.

Division by $n-2$ is a "degrees of freedom" correction, because two parameters ($\beta_0$ and $\beta_1$) have been estimated.

When $n$ is large, it doesn't matter whether $n$, $n-1$, or $n-2$ are used.

### 1.3.3 Adjusted $R^2$

The measure $R^2$ defined earlier keeps on increasing as we add extra explanatory variables and thus **not take account of the degrees of freedom problem**.

*增加变量会增强模型的拟合能力，RSS会相应减小，$R^2 = 1-\frac{\mathrm{RSS}}{\mathrm{TSS}}$则增大，直到等于1. 过度增加变量会导致过拟合。*

The adjusted $R^2$ is simply $R^2$ adjusted for degrees of freedom.

$$
1-\bar{R}^2 = \frac{n-1}{n-(k+1)}(1-R^2) \tag{7}
$$

where $k$ is the number of regressors.

*参数比变量多一个$\beta_0$.*

If $R^2$ does not increase significantly on the addition of a new independent variable, then the value of $\bar{R}^2$ will actually decrease. Vice versa.

## 1.4 The Least Square Assumption for Causal Inference

We have treated OLS as a way to draw a straight line through the data on $Y$ and $X$. We want to know under what conditions does the slope of this line have a causal interpretation?

**The least square assumption for causal inference**:

1. The conditional distribution of $u$ given $X$ has mean zero, that is $E(u|X=x)=0$
   - It implies that $X_i$ and $u_i$ are uncorrelated. *这就意味着X是一个足够独立的变量在影响Y，而不会通过u作用于Y。*
2. $(X_i, Y_i)$ are independently and indentically distributed.
3. Large outliers in $X$ and/or $Y$ are rare.