EC508: Econometrics
Mechanics of OLS

Jean-Jacques Forneron

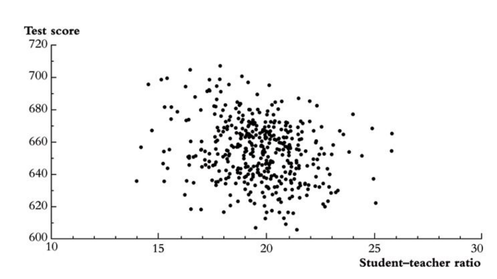
Spring, 2023

Boston University

Mechanics of OLS

- The population regression line: Test Score $= \beta_0 + \beta_1$STR

$$\beta_1 = \frac{\Delta \text{Test Score}}{\Delta \text{STR}} = ???$$



- Finite Sample: only observed n data points, not the population regression line β_1

OLS Estimator: $\min_{b_0,b_1} \sum_{i=1}^{n}(Y_i - [b_0 + b_1 X_i])^2$

- The OLS estimator minimizes the average squared difference between the actual values of Y_i and the prediction ("predicted value") based on the estimated line.
- This minimization problem can be solved using calculus
- The result is the OLS estimators of β_0 and β_1

OLS Estimator: Deriving the Estimator

- The OLS estimator minimizes

$$\min_{b_0, b_1} \sum_{i=1}^{n} (Y_i - [b_0 + b_1 X_i])^2$$

- First order conditions:

$/b_0 : \sum_{i=1}^{n} (Y_i - [\hat{\beta}_0 + \hat{\beta}_1 X_i]) = 0$
$/b_1 : \sum_{i=1}^{n} X_i (Y_i - [\hat{\beta}_0 + \hat{\beta}_1 X_i]) = 0$

- Imply:

$$\begin{pmatrix} 1 & \frac{1}{n}\sum_{i=1}^{n} X_i \\ \frac{1}{n}\sum_{i=1}^{n} X_i & \frac{1}{n}\sum_{i=1}^{n} X_i^2 \end{pmatrix} \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} = \begin{pmatrix} \frac{1}{n}\sum_{i=1}^{n} Y_i \\ \frac{1}{n}\sum_{i=1}^{n} Y_i X_i \end{pmatrix}$$

- And Finally:

$$\hat{\beta}_0 = \bar{Y}_n - \hat{\beta}_1 \bar{X}_n, \quad \hat{\beta}_1 = \frac{\frac{1}{n}\sum_{i=1}^{n} X_i [Y_i - \bar{Y}_n]}{\frac{1}{n}\sum_{i=1}^{n} X_i [X_i - \bar{X}_n]}$$

The OLS Estimator, Predicted Values, and Residuals

KEY CONCEPT 4.2

The OLS estimators of the slope β_1 and the intercept β_0 are

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2} = \frac{s_{XY}}{s_X^2} \tag{4.7}$$

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}. \tag{4.8}$$

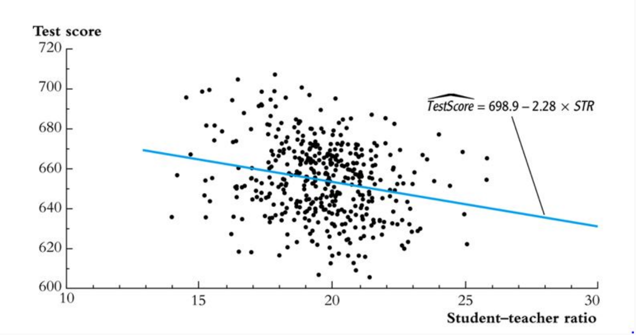
The OLS predicted values $\hat{Y}_i$ and residuals $\hat{u}_i$ are

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i,\ i = 1, \ldots, n \tag{4.9}$$

$$\hat{u}_i = Y_i - \hat{Y}_i,\ i = 1, \ldots, n. \tag{4.10}$$

The estimated intercept $(\hat{\beta}_0)$, slope $(\hat{\beta}_1)$, and residual $(\hat{u}_i)$ are computed from a sample of n observations of X_i and Y_i, $i = 1, \ldots, n$. These are estimates of the unknown true population intercept (β_0), slope (β_1), and error term (u_i).

Application to the California Test Score – Class Size data



- Estimated slope $\hat{\beta}_1 = -2.28$
- Estimated intercept $\hat{\beta}_0 = 698.9$
- Estimated regression line: Test Score $= 698.9 – 2.28 \times$ STR

Application to the California Test Score in R

```r
# package to open data set
library(foreign)

# open data set
data = read.dta('caschool.dta')

# construct score variable
data$score = 0.5*(data$math_scr + data$read_scr)

# run the ols regression
lm(score~str,data=data)
```

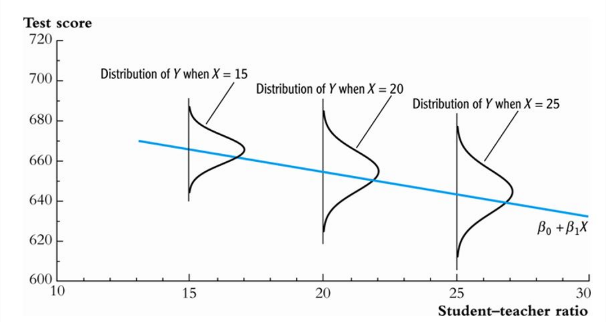
The Least Squares Assumptions (SW 4.4)

- What, in a precise sense, are the properties of the sampling distribution of the OLS estimator? When will be unbiased? What is its variance?
- To answer these questions, we need to make some assumptions about how Y and X are related to each other, and about how they are collected (the sampling scheme)
- These assumptions – there are three – are known as the Least Squares Assumptions.

The Least Squares Assumptions

1. The conditional distribution of u given X has mean zero, that is, $\mathbb{E}(u|X = x) = 0$
   $\Rightarrow$ This implies that $\hat{\beta}_1$ is unbiased
2. $(X_i, Y_i), i = 1, \ldots, n$, are i.i.d.
   - This is true if (X, Y) are collected by simple random sampling
   - This delivers the sampling distribution of $\hat{\beta}_0$ and $\hat{\beta}_1$
3. Large outliers in X and/or Y are rare.
   - Technically, X and Y have finite fourth moments
   - Outliers can result in meaningless values of β_1

Least squares assumption #1: $\mathbb{E}(u|X = x) = 0$



Example: Test Score$_i = \beta_0 + \beta_1 \text{STR}_i + u_i$, u_i = unobserved factors

- What are some of these "other factors"?
- Is $\mathbb{E}(u|X = x) = 0$ plausible for these other factors?

Least squares assumption #1 cont'd

- A benchmark for thinking about this assumption is to consider an ideal randomized controlled experiment:
- X is randomly assigned to people (students randomly assigned to different size classes; patients randomly assigned to medical treatments). Randomization is done by computer – using no information about the individual.
- Because X is assigned randomly, all other individual characteristics – the things that make up u – are distributed independently of X, so u and X are independent
- Thus, in an ideal randomized controlled experiment, $\mathbb{E}(u|X = x) = 0$ (that is, LSA #1 holds)
- In actual experiments, or with observational data, we will need to think hard about whether $\mathbb{E}(u|X = x) = 0$ holds.

Least squares assumption #2 $(X_i, Y_i), i = 1, \ldots, n$, are i.i.d.

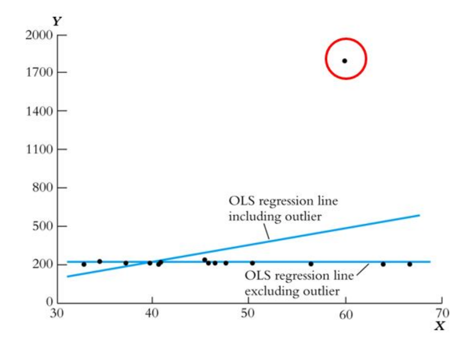
- This arises automatically if the entity (individual, district) is sampled by simple random sampling:
  - The entities are selected from the same population, so (X_i, Y_i) are identically distributed for all $i = 1, \ldots, n$.
  - The entities are selected at random, so the values of (X, Y) for different entities are independently distributed.
- The main place we will encounter non-i.i.d. sampling is when data are recorded over time for the same entity (panel data and time series data) – we will deal with that complication when we cover time-series.

Least squares assumption #3 Large outliers are rare

Technical Statement: $\mathbb{E}(X^4) < \infty$ and $\mathbb{E}(Y^4) < \infty$

- A large outlier is an extreme value of X or Y
- On a technical level, if X and Y are bounded, then they have finite fourth moments. (Standardized test scores automatically satisfy this; STR, family income, etc. satisfy this too.)
- The substance of this assumption is that a large outlier can strongly influence the results – so we need to rule out large outliers.
- Look at your data! If you have a large outlier, is it a typo? Does it belong in your data set? Why is it an outlier?

OLS can be sensitive to an outlier



- Is the lone point an outlier in X or Y?
- In practice, outliers are often data glitches (coding or recording problems). Sometimes they are observations that really shouldn't be in your data set. Plot your data!