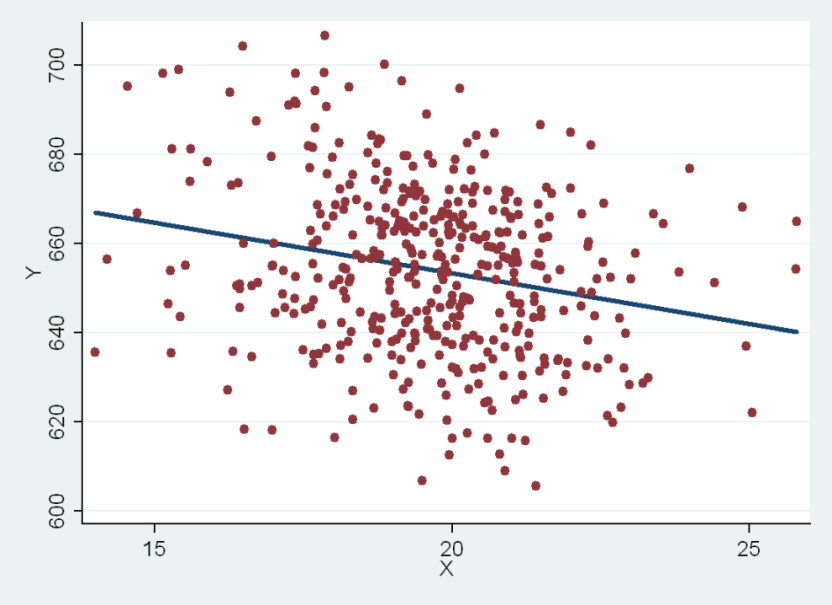
EC 3303: Econometrics I

Linear Regression with One Regressor (Part 2)



Kelvin Seah

AY 2022/2023, Semester 2

Measures of Fit

A natural question is how well the OLS regression line **"fits"** or **describes** the data. Two regression statistics provide complementary measures of the quality of fit:

- R^2 : measures the fraction of the variance of Y that is explained by X; it is ***unitless*** and ranges between zero (no fit) and one (perfect fit).

- ***Standard error of the regression (SER)*** : measures the magnitude of a typical regression residual ***in units of Y***.

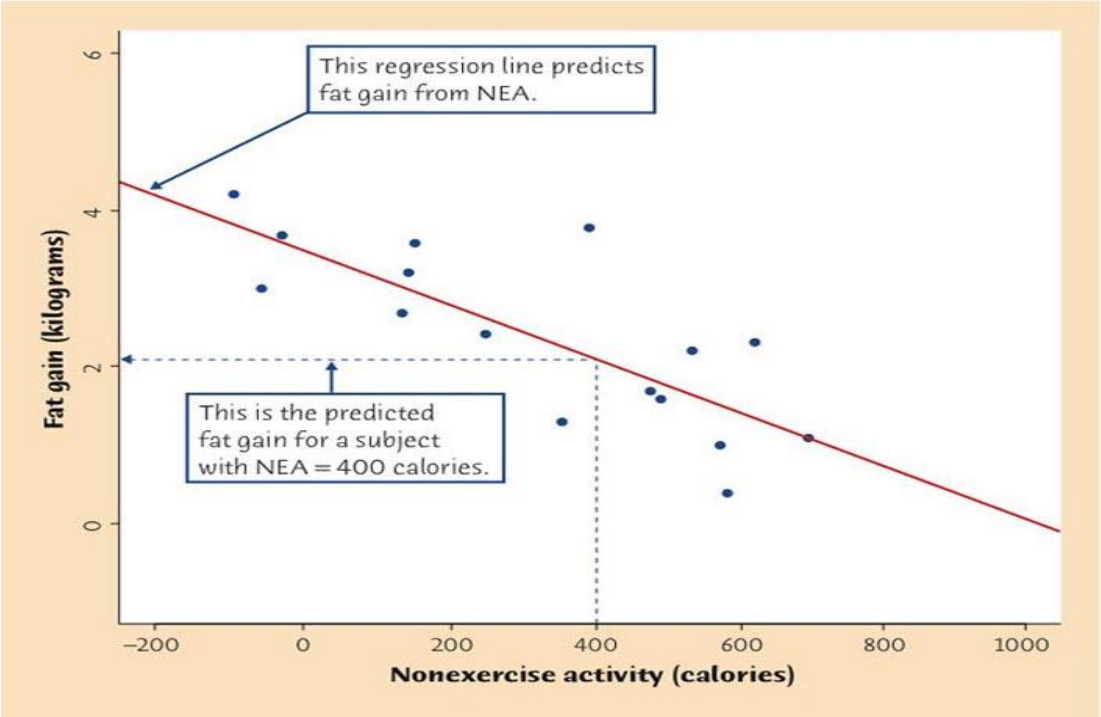
R^2

- R^2: fraction of the variation in the values of Y that is explained by the OLS regression of Y on X.

- In the case of a single regressor, R^2 is the square of the sample correlation between X and Y.

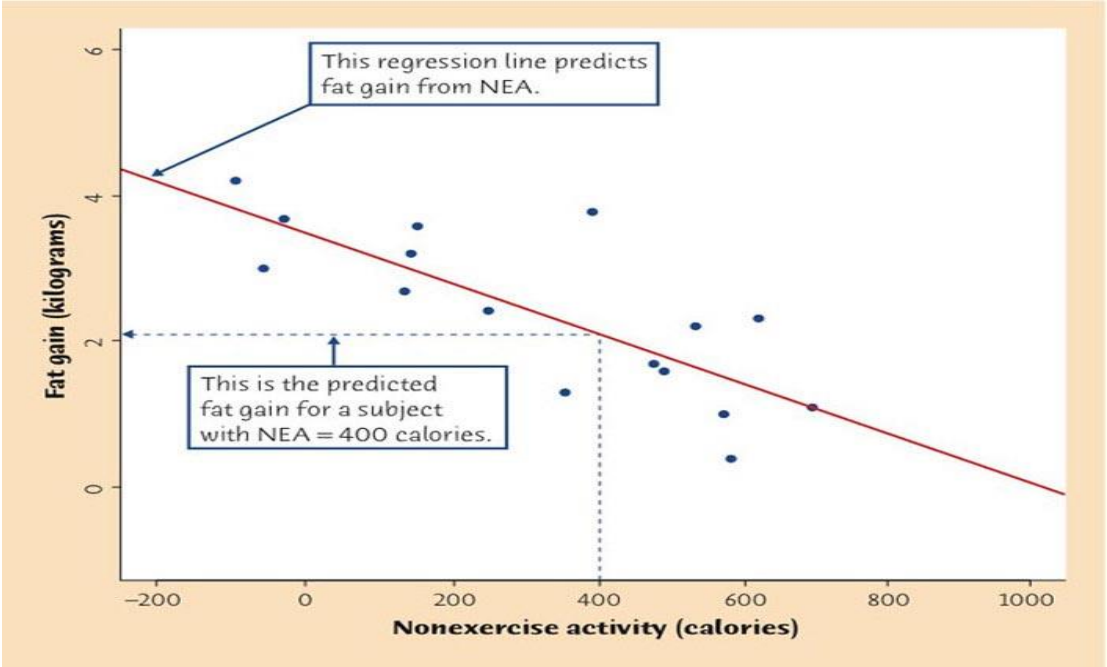
  - Perfect correlation ($R = 1$ *or* $R = -1$) means all the points lie exactly on a line. $R^2 = 1$ and all of the variation in Y is accounted for by the linear relationship with X.

- R^2 is based on the fact that there are ***two sources*** of variation in Y: an ***explained*** source and an ***unexplained*** source.



- One reason the fat gains of the individuals vary is that there is a relationship between fat gain (Y) and activity level (X).
- As X increases from -94 calories to 690 calories among the 16 individuals, it "pulls" fat gain Y with it along the regression line.
- The OLS regression explains this part of the variation in fat gains.

- The fat gains do not lie exactly on the OLS regression line but are scattered above and below it.
- This is the second source of variation in Y, the part that is unexplained.



- If $R^2 = 0.61$. 61% of the variation in fat gains is ***explained*** by the straight line relationship between fat gain and activity (or 61% of the variation in fat gains is explained by activity, in this one regressor case) . The remaining 39% of the variation (depicted by the vertical scatter) is ***unexplained***.

R^2 More Technically

$$Y_i = \hat{Y}_i + \hat{u}_i \qquad (6)$$

- R^2: ratio of the sample variance of $\hat{Y}_i$ to the sample variance of Y_i

$$R^2 = \frac{ESS}{TSS} = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \bar{\hat{Y}})^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} \qquad (7)$$

- R^2 can also be written in terms of the fraction of the variance of Y_i not explained by the regression.

$$R^2 = 1 - \frac{SSR}{TSS} = 1 - \frac{\sum_{i=1}^{n}\hat{u}_i^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} \qquad (8)$$

Total Sum of Squares =

Explained Sum of Squares + Sum of Squared Residuals

- $0 \leq R^2 \leq 1$;
  - If $R^2 = 0$, the regression (or X_i) explains none of the variation in Y_i and $ESS = 0$.
  - If $R^2 = 1$, the regression explains all of the variation in Y_i and $ESS = TSS$.
  - An R^2 near 1 indicates that the regression (or X_i) is good at predicting Y_i . An R^2 near 0 indicates that the regression is not good at predicting Y_i .

Example: OLS regression – Stata output

```
regress testscr str, robust

Regression with robust standard errors                Number of obs =     420
                                                      F(  1,   418) =   19.26
                                                      Prob > F      =  0.0000
                                                      R-squared     =  0.0512
                                                      Root MSE      =  18.581

------------------------------------------------------------------------------
             |               Robust
    testscr  |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         str |  -2.279808   .5194892    -4.39   0.000    -3.300945   -1.258671
       _cons |    698.933   10.36436    67.44   0.000     678.5602    719.3057
------------------------------------------------------------------------------
```

- R^2 =0.051 means that the regression (or the regressor, STR, in this one regressor case) explains 5.1% of the variance of TestScore; STR does explain some of the variation in TestScore, but much variation remains unaccounted for.

Standard Error of the Regression (SER)

- The *SER* is a measure of the spread of the observations around the regression line (measured in units of Y).
- More precisely, *SER* is the sample standard deviation of the OLS residuals:

$$SER = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(\hat{u}_i - \bar{\hat{u}})^2} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}\hat{u}_i^2}$$

where $\frac{1}{n-2}\sum_{i=1}^{n}(\hat{u}_i - \bar{\hat{u}})^2$ is the sample variance of the OLS residuals.

[note: second equality holds because $\bar{\hat{u}} = \frac{1}{n}\sum_{i=1}^{n}\hat{u}_i = 0$)].

Root Mean Squared Error

Root mean squared error (*RMSE*) is closely related to the *SER*:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\hat{u}_i^2}$$

measures same thing as the *SER* – the minor difference is the division by n instead of $n–2$.

[note: when n is large, it makes negligible difference whether n or n-*2* are used]

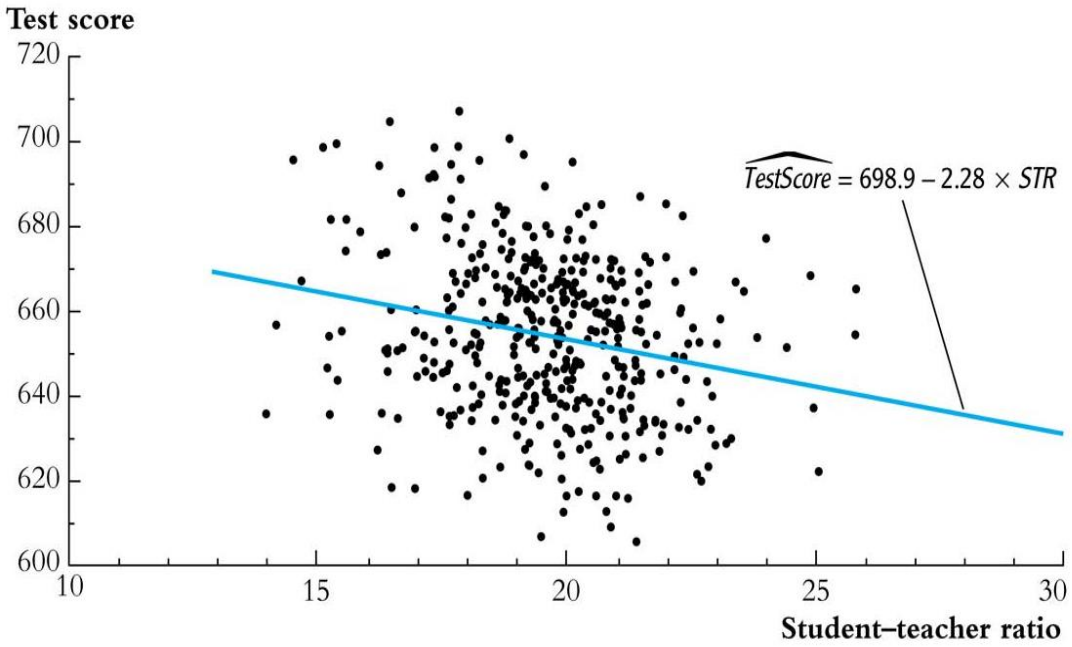
Example: OLS Regression – Stata output

```
regress testscr str, robust

Regression with robust standard errors                Number of obs =     420
                                                      F(  1,   418) =   19.26
                                                      Prob > F      =  0.0000
                                                      R-squared     =  0.0512
                                                      Root MSE      =  18.581

------------------------------------------------------------------------------
             |               Robust
    testscr  |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         str |  -2.279808   .5194892    -4.39   0.000    -3.300945   -1.258671
       _cons |    698.933   10.36436    67.44   0.000     678.5602    719.3057
------------------------------------------------------------------------------
```

- SER of 18.6 means that the sample standard deviation of the OLS residuals is 18.6.
- There is a large spread of the scatterplot around the regression line.
- Large spread means that predictions of test scores using the OLS regression line will often be wrong by a large amount.



$R^2 = .05$, $SER = 18.6$

What to make of the low R^2 and large SER?

- This does not imply that this regression is either "good" or "bad".
- What low R^2 does say is that STR alone explains only a small part of the variation in scores. Test scores are influenced by other important factors as well.

Unit Change

$$\widehat{Y_i} = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

- Dependent variable: from Y to aY

$$\widehat{aY_i} = a\hat{\beta}_0 + a\hat{\beta}_1 X_i$$

E.g.: $$\widehat{Earn} = 239.16 + 5.20 \times Age$$

$Earn$ denotes earnings in S$ and Age denotes age in years.

How would the estimated coefficients change if earnings were converted to Ringgit (1S$=3Ringgit)?

$$\widehat{Earn} = (3 \times 239.16) + (3 \times 5.20) \times Age$$

Unit Change

$$\widehat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

- Regressor: from X to bX

$$\widehat{Y}_i = \hat{\beta}_0 + \frac{\hat{\beta}_1}{b} (bX_i)$$

E.g.:

$$\widehat{Earn} = 239.16 + 5.20 \times Age$$

$Earn$ denotes earnings in S$ and Age denotes age in years.

How would the estimated coefficients change if age were converted to months?

$$\widehat{Earn} = 239.16 + \frac{5.20}{12} \times (12 \times Age)$$

Least Squares Assumptions

- What are the conditions under which OLS will give appropriate estimates of the population coefficients β_0 and β_1?

- There are 3 conditions – known as the "Least Squares Assumptions".

Least Squares Assumption #1

Population regression model:

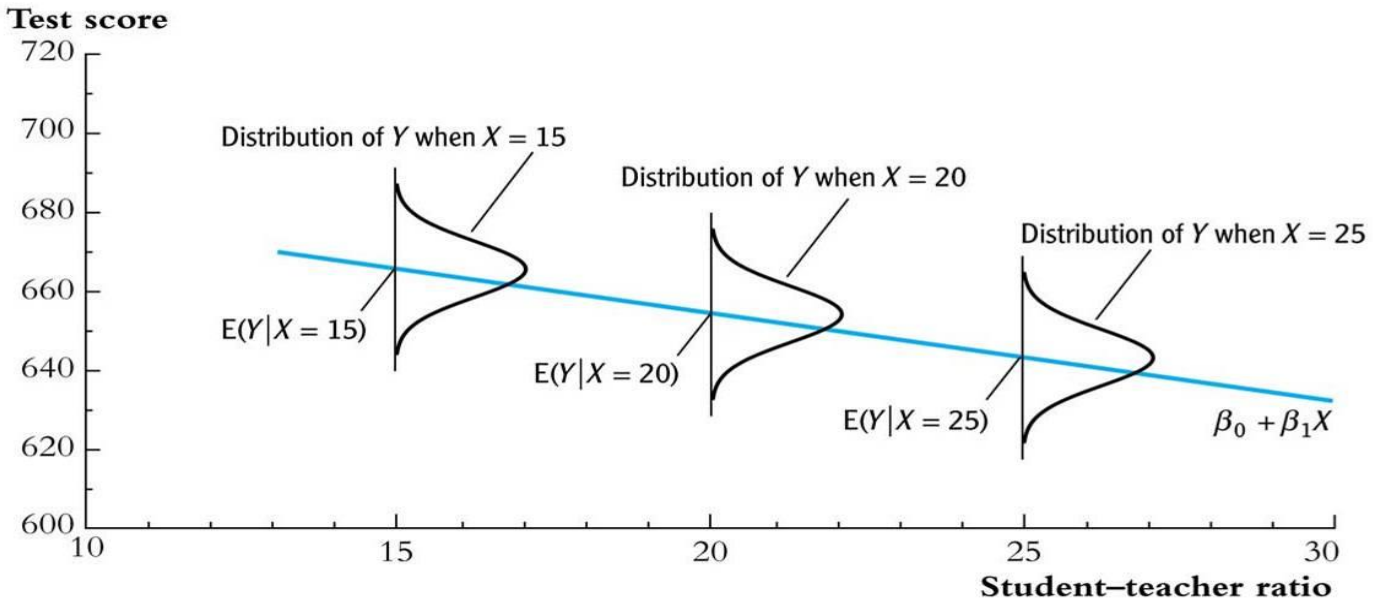
$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

- LSA #1: given X_i, the conditional distribution of the population error term u_i, has a mean of zero.

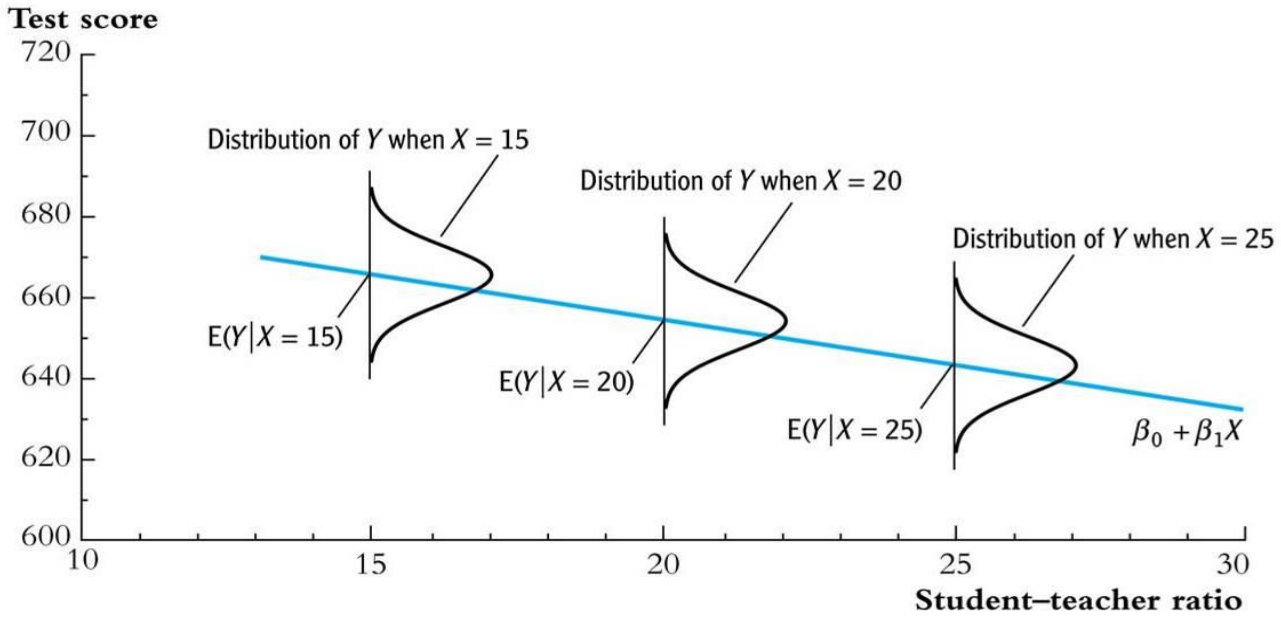
$$E(u_i | X_i = x) = 0$$

  - implies the other factors not explicitly included in the model (therefore subsumed in u_i) are uncorrelated with X_i.

Least Squares Assumption #1



- At a given value of X, Y is distributed around the regression line.
- The error $u = Y - (\beta_0 + \beta_1 X)$ has a conditional mean of zero, for all values of X.



- At a given class size, say 20, sometimes the "other factors" lead to better performance than predicted by the population regression line ($u_i > 0$) and sometimes to worse performance ($u_i < 0$); However, ***on average***, the prediction is correct (i.e: $E(u_i|X_i = 20) = 0$).

- This applies to any other value of class size as well.

Assumption

$$E(u_i|X_i = x) = 0$$

implies that

$$Cov(u_i, X_i) = 0$$

- arises because, if the conditional mean of u does not depend on X, then u and X are uncorrelated.

If the assumption

$$E(u_i|X_i = x) = 0$$

holds,

$$E(\hat{\beta}_1) = \beta_1 \text{ and } E(\hat{\beta}_0) = \beta_0$$

LSA #1:

$$E(u_i|X_i = x) = 0$$

In class size e.g.,

$$TestScore_i = \beta_0 + \beta_1 STR_i + u_i, \text{ where } u_i\text{=other factors}$$

- what are some of these "other factors"?
- Is $E(u_i|X_i = x) = 0$ plausible for these other factors?
  - whether the assumption holds in a given application requires careful thought.

Least Squares Assumption #1

A benchmark for thinking about this assumption is to consider a ***randomized controlled experiment***:

- X (i.e. treatment) is randomly assigned.
- Because X is assigned randomly, all other individual characteristics – the things that make up u – are independently distributed of X.
- Thus, in a randomized controlled experiment,

$$E(u_i|X_i = x) = 0$$

- With observational data, need to think hard about whether $E(u_i|X_i = x) = 0$ holds.

Least Squares Assumption #2

- LSA #2: (X_i, Y_i), $i = 1, \ldots, n$ are independently and identically distributed (i.i.d).

- arises automatically if the entity (individual, district) is randomly selected by simple random sampling.

- main case where we encounter non-i.i.d. sampling is when data are recorded over time ("time series data").

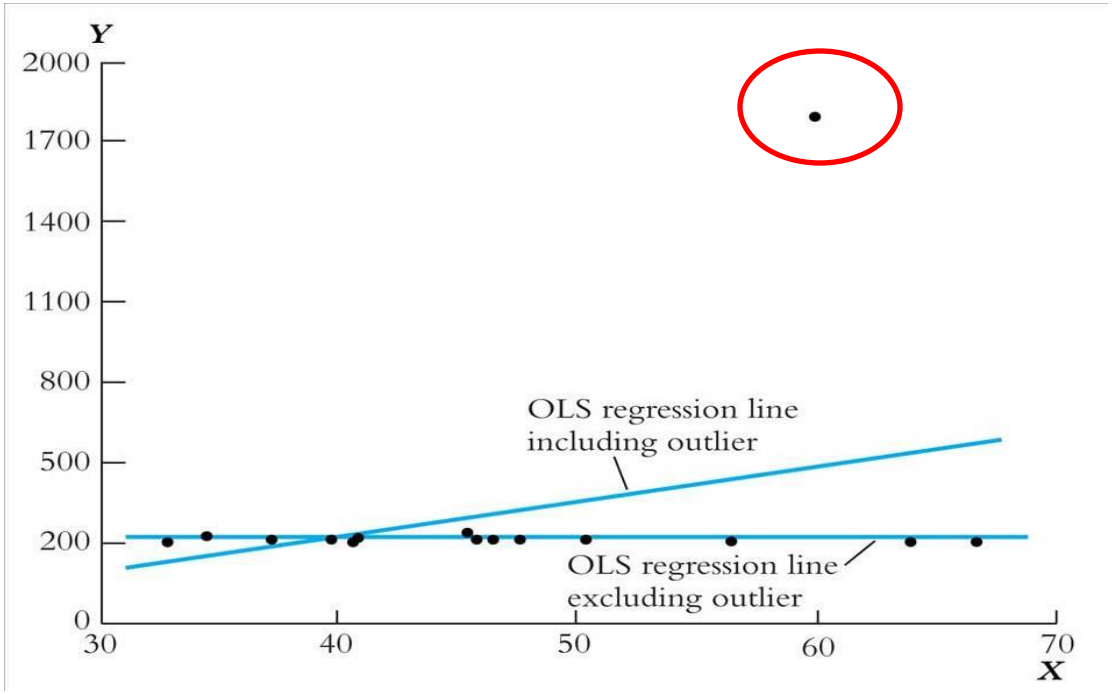
  - e.g. immigration rates and GDP from one country over time.

  - time series data violates the independence part of the i.i.d assumption.

Least Squares Assumption #3

- LSA#3: Large outliers are rare
  - outliers are observations with values of either X_i or Y_i, or both, that are far outside the usual range.
  - outliers can strongly influence the OLS regression results and give misleading answers.

OLS can be sensitive to outliers



- Outliers often are data glitches (coding/recording problems) – so check your data for outliers!
- Produce a scatterplot.

Sampling Distribution of the OLS Estimators

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

- $\hat{\beta}_1$ & $\hat{\beta}_0$ are calculated from a sample of data; a different sample will give a different value of $\hat{\beta}_1$ and $\hat{\beta}_0$.
- OLS estimators are random variables with a sampling distribution.
- Focus on the sampling distribution of $\hat{\beta}_1$.

We want to:

- describe the "sampling distribution" of $\hat{\beta}_1$
- use $\hat{\beta}_1$ to test hypotheses such as $\beta_1 = 0$
- construct a confidence interval for β_1

Sampling Distribution of $\hat{\beta}_1$

Questions to address:

- What is $E(\hat{\beta}_1)$? (where is it centered?)
- What is var($\hat{\beta}_1$)? (measure of sampling uncertainty)
- What is the distribution of $\hat{\beta}_1$ in small samples?
  - Answer: It can be very complicated in general.
- What is the distribution of $\hat{\beta}_1$ in large samples?
  - Answer: In large samples, $\hat{\beta}_1$ is ***normally*** distributed.

Properties of OLS estimators under 3 LSA…

If the 3 Least Squares Assumptions hold, then

- the exact (finite sample) sampling distribution of $\hat{\beta}_1$ is such that:

$$E(\hat{\beta}_1) = \beta_1$$

- $\hat{\beta}_1$ is an unbiased estimator of the population slope β_1.

$$Var(\hat{\beta}_1) = \frac{1}{n} \times \frac{Var[(X_i - \mu_X)u_i]}{[Var(X_i)]^2}$$

- Apart from its mean and variance, the exact distribution of $\hat{\beta}_1$ is complicated.

- If n is sufficiently large, however, then by the Central Limit Theorem, $\hat{\beta}_1$ has an approximate normal distribution.

As $n \to \infty$,

$$\hat{\beta}_1 \xrightarrow{d} N\left(\beta_1, \frac{1}{n}\frac{Var[(X_i - \mu_X)u_i]}{[Var(X_i)]^2}\right)$$

- When n is large, the distribution of the standardized $\hat{\beta}_1$, $\frac{\hat{\beta}_1 - E(\hat{\beta}_1)}{\sqrt{Var(\hat{\beta}_1)}}$, is well approximated by a $N(0,1)$ distribution:

As $n \to \infty$

$$\frac{\hat{\beta}_1 - E(\hat{\beta}_1)}{\sqrt{Var(\hat{\beta}_1)}} \xrightarrow{d} N(0,1)$$

When n is large,

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{1}{n}\frac{Var[(X_i - \mu_X)u_i]}{[Var(X_i)]^2}\right)$$

- How large must n be?
  - $n \geq 100$ is sufficiently large, in most cases.

Implications of Sampling Distribution of $\hat{\beta}_1$

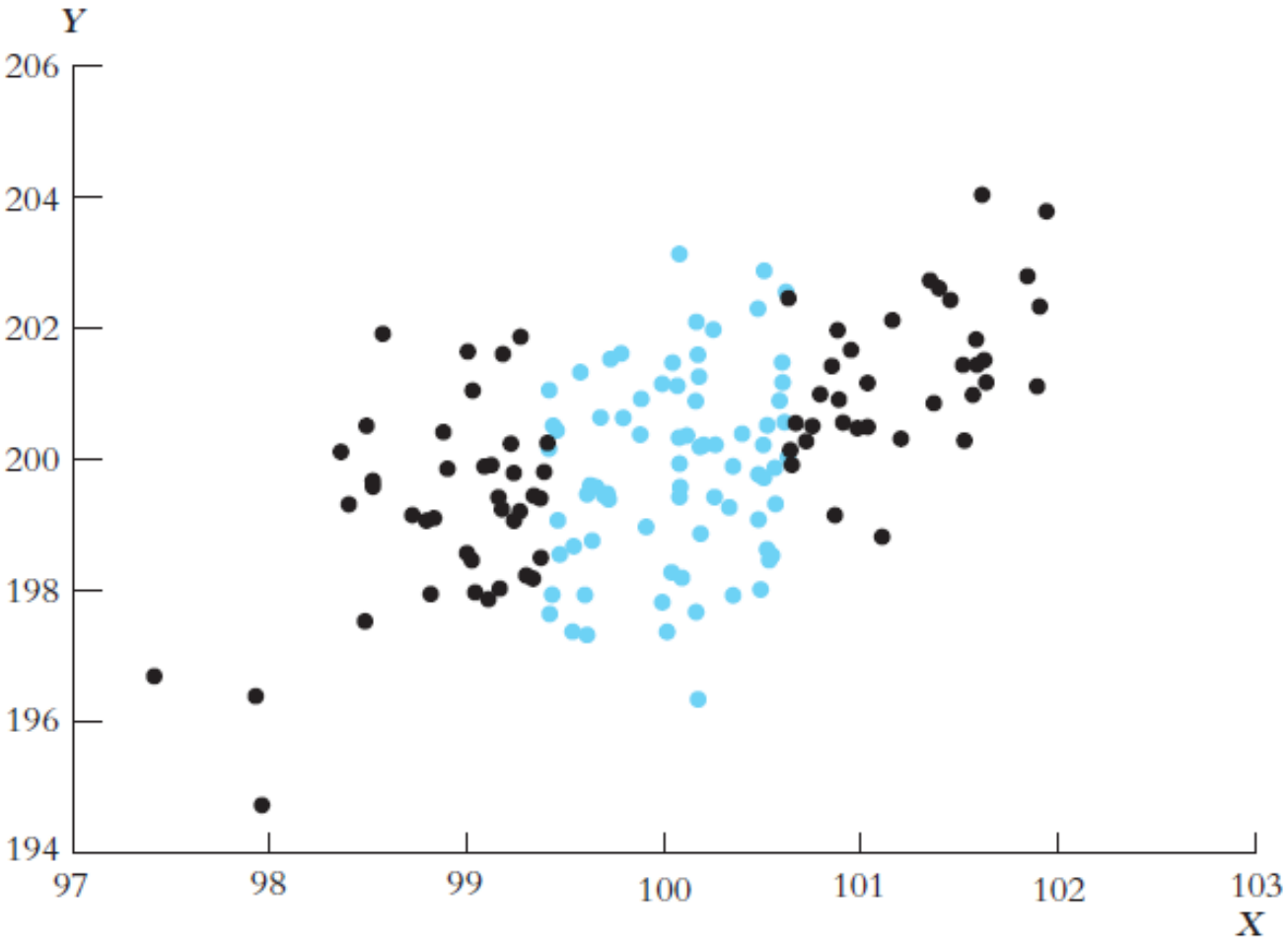
(1) Because $Var(\hat{\beta}_1) \propto \frac{1}{n}$ and $E(\hat{\beta}_1) = \beta_1$, $\hat{\beta}_1$ is a consistent estimator of β_1

$$\hat{\beta}_1 \xrightarrow{p} \beta_1$$

When the sample size n is large, $\hat{\beta}_1$ will be close to β_1 with high probability.

(2) Because $Var(\hat{\beta}_1) = \frac{1}{n} \times \frac{Var[(X_i - \mu_X)u_i]}{[Var(X_i)]^2}$, the larger the variance of X_i, the smaller the variance of $\hat{\beta}_1$. In other words, the larger the variance of X_i, the more precise is $\hat{\beta}_1$.

We are now ready to turn to hypothesis tests & confidence intervals…


Y
206
204
202
200
198
196
194
97
98
99
100
101
102
103
X

Plan for Next Two Lectures

- As mentioned in the course syllabus, as we are exploring the use of different modes of lecture presentation for subsequent semesters, for next week's lecture (Week 4) and the lecture in Week 5, instead of the usual recorded lecture, we will use a recorded webcast.
- Both lectures will be uploaded on Canvas "Videos/Panopto" as usual.
- After the lectures, a poll will then be conducted to explore which mode of presentation students prefer.

Announcement

- Please be reminded that tutorials begin in week 3 (i.e. week starting 23 Jan) for students in the **TD** groups.

- 23 January (Monday) and 24 January (Tuesday) are public holidays. If your tutorial class happens to fall on these days, please attend any other tutorial slot in week 3 or week 4 as a make-up (below, lists the possible tutorial slots you can attend as a make-up). Please keep your assigned tutor informed of your make-up tutorial attendance and participation when you meet your tutor during your second tutorial.

Possible Make-up Classes

Tutorial Class	Day	Start	End	Venue	Odd/Even
TD3	Wed	10:00	12:00	AS4-0110	Odd Week
TD4	Wed	14:00	16:00	AS4-0110	Odd Week
TE2	Tue	10:00	12:00	AS4-0110	Even Week
TE4	Thu	10:00	12:00	AS4-0110	Even Week
TE5	Thu	14:00	16:00	AS4-0110	Even Week

Announcement

- Tutorials will begin in week 4, (i.e. week starting 30 Jan) for students in the **TE** groups.