Topic 4: Simple Linear Regression-Estimation

Part B

3) Fit

- So how well does the estimated model explain the dependent variable?
- R^2 is the fraction of the variation of Y that is explained by the model
- Here model is linear functional form and one explanatory variable X
- The dependent variable Y varies from observation to observation, and so we just want to know how much of that variation we have captured with our model.

3) Fit

- Explained Sum of Squares $ESS = \sum(\hat{Y}_i - \bar{Y})^2$
- Sum of Squared Residuals $SSR = \sum(Y_i - \hat{Y}_i)^2$
- Total Sum of Squares $TSS = \sum(Y_i - \bar{Y})^2$
- We can show $ESS + SSR = TSS$
- What does this mean?

$Y_i = \hat{Y}_i + \hat{u}_i =$ OLS prediction $+$ OLS residual

$\Rightarrow$sample variance$(Y_i) =$ sample variance$(\hat{Y}_i) +$ sample variance$(\hat{u}_i)$

$\Rightarrow$total sum of squares (TSS: total variation) $=$

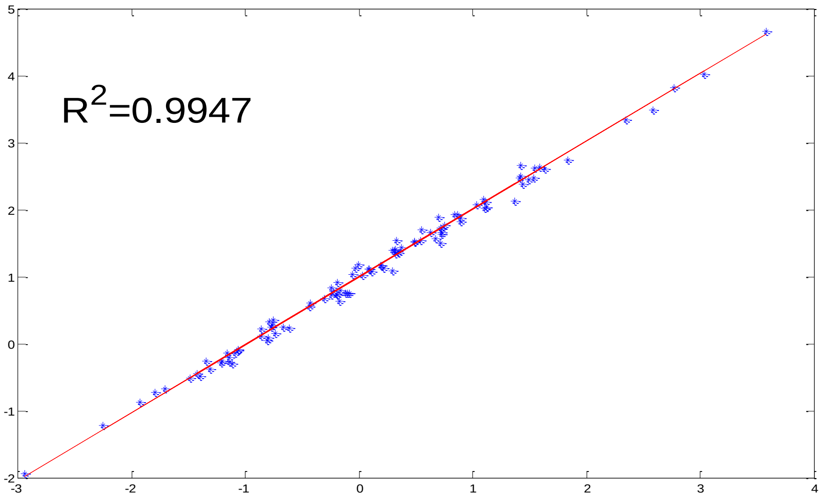
"explained" sum of squares (ESS: variation explained by OLS)

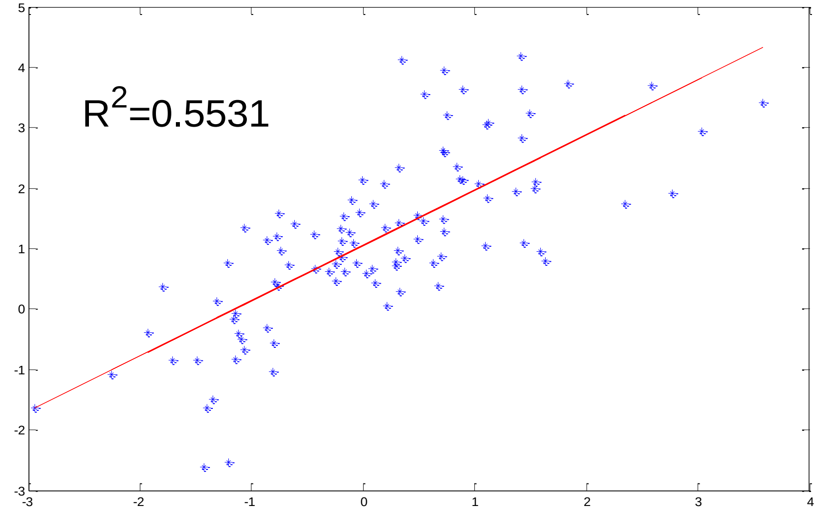
$+$"residual" sum of squares (SSR: variation that cannot be explained by OLS)

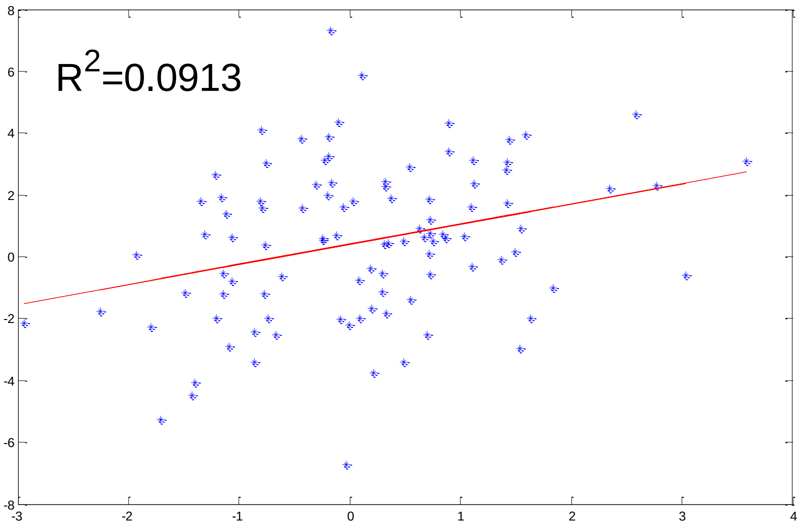
3) Fit

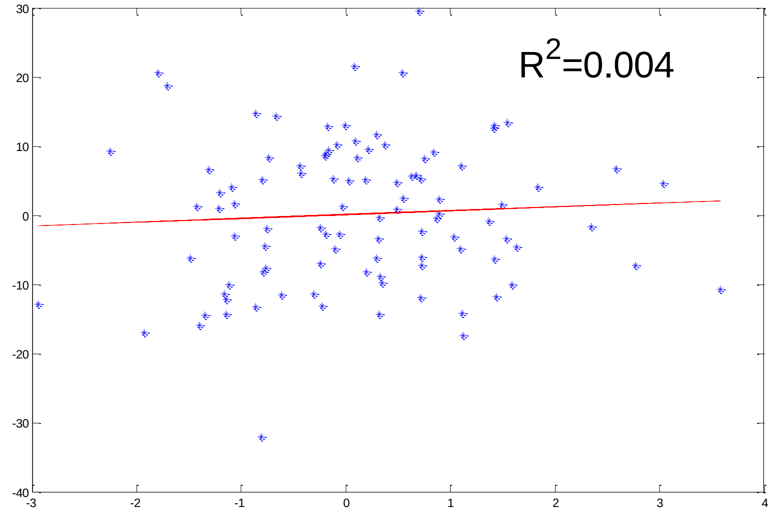
- $R^2 = \dfrac{ESS}{TSS}$ or $R^2 = 1 - \dfrac{SSR}{TSS}$
- $0 \leq R^2 \leq 1$
- R^2 near zero means model explained virtually none of the variation of Y_i
- R^2 near one means model explained virtually all of the variation of Y_i.

3) Fit









3) Fit

- The numerical example: recall:

X_i	Y_i
1	2
2	5
3	5

$\Rightarrow$

$\hat{Y}_i$	$\hat{u}_i$	$\hat{u}_i^2$	$(Y_i - \bar{Y})^2$
2.5	-0.5	0.25	4
4	1	1	1
5.5	-0.5	0.25	1

- Recall: $\hat{Y}_i = 1 + 1.5X_i$

$$R^2 = 1 - \frac{SSR}{TSS} = 1 - \frac{\sum \hat{u}_i^2}{\sum (Y_i - \bar{Y})^2} = 1 - \frac{1.5}{6} = 0.75$$

3) Fit

- Is $R^2=0.75$ good? bad?
- Shouldn't think of R^2 as good or bad.
- If it's low, e.g. $R^2=0.05$ Why?
  - A lot of things could be causing Y other than just X. So a low R^2 could mean we need to include other explanatory variables.

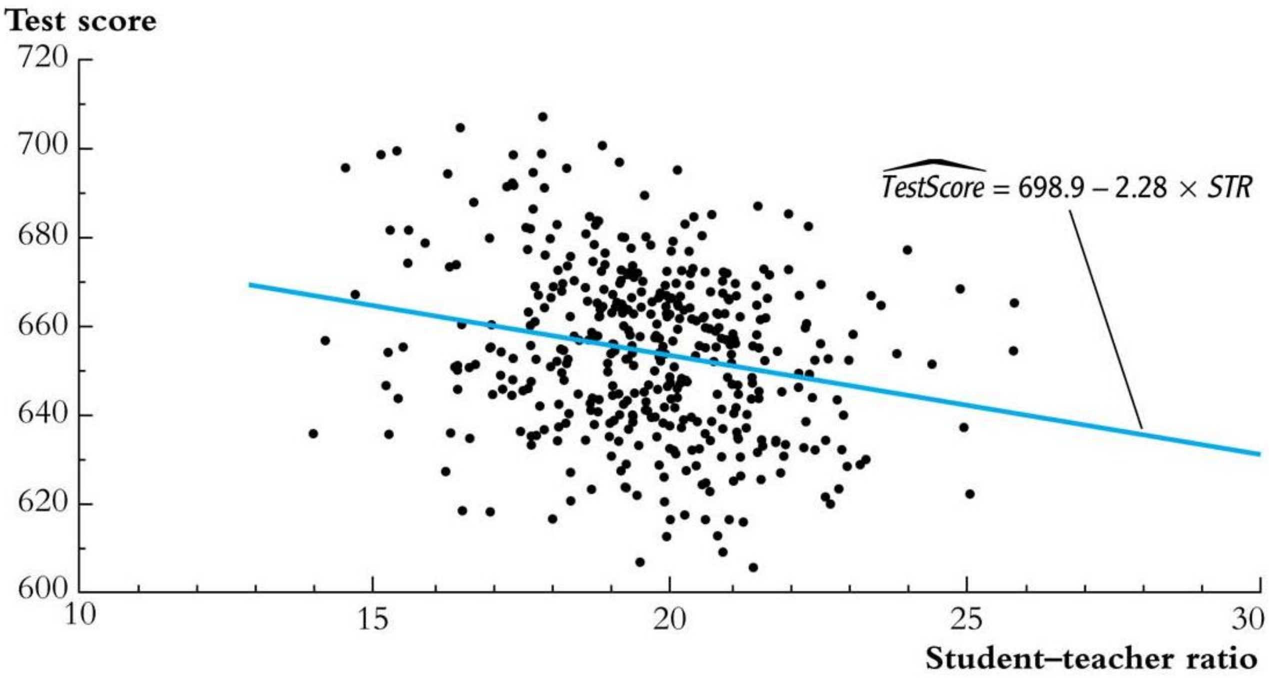
3) Fit

- Standard error of regression (SER)
- The SER is an estimator of the standard deviation of the error u_i. It is a measure of spread (just like standard deviation).

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

- Note that $\bar{\hat{u}} = \frac{1}{n}\sum \hat{u}_i = 0$.
- Here it is divided by n-2, since we "used" up to 2 degree of freedom by estimating beta0 and beta1.
- SER measures the average "size" of the OLS residual (the average "mistake" made by the OLS regression line)

- Example of the R^2 and SER



$$\widehat{TestScore} = 698.9 - 2.28 \times STR, \ \mathbf{R^2 = .05,\ SER = 18.6}$$

4) OLS Assumptions

- Why do we use the OLS estimator?
- What are the properties of the OLS estimator?
- Under what conditions is the OLS estimator "nice"?

  Unbiased, consistent, distributed normally, efficient.

- 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 are known as the Least Squares Assumptions.

4) OLS Assumptions

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

Assumption 1: The conditional distribution of u_i given X_i has mean zero, that is, $E(u_i \mid X_i) = 0$.

- *This is the key assumption to ensure that* $\hat{\beta}_1$ *is unbiased. It may not be satisfied in practice.*

Assumption 2: (X_i, Y_i), $i = 1,\ldots,n$, are i.i.d.

- *This is true if* (X_i, Y_i) *are collected by simple random sampling*

Assumption 3: Large outliers in X_i and Y_i are rare.

- *Technically,* X_i *and* Y_i *have finite fourth moments*
- *Outliers can result in meaningless values of* $\hat{\beta}_1$

4) OLS Assumptions

- Assumption 1: $E(u_i|X_i) = 0$.
- Given the explanatory variable X_i, the expected value of the error is zeros.
- This implies that X_i and u_i are not correlated.

$$E(u_i|X_i) = 0 \Rightarrow E(u_i) = E[E(u_i|X_i)] = 0 \text{ by law of iterated expectation}$$

$$\begin{aligned} cov(X_i, u_i) &= E(X_i u_i) - E(X_i)E(u_i) = E(X_i u_i) \\ &= E[E(X_i u_i|X_i)] \text{ by law of iterated expectation} \\ &= E\left[X_i \cdot \underbrace{E(u_i|X_i)}_{=0} \right] = 0 \end{aligned}$$

- If X_i and u_i are not correlated, this does NOT imply $E(u_i|X_i)=0$
- If X_i and u_i are correlated, we know that Assumption 1 is false.

4) OLS Assumptions

- **Assumption 1**: $E(u_i|X_i) = 0$.
- A benchmark for thinking about this assumption is to consider an ideal randomized controlled experiment:
- X_i is randomly assigned to people. Randomization is done by computer – using no information about the individual.
- Because X_i is assigned randomly, all other individual characteristics – the things that make up u_i – are independently distributed of X_i
- Thus, in an ideal randomized controlled experiment, $E(u_i|X_i) = 0$ (that is, Assumption #1 holds)
- With observational data, we will need to think hard about whether $E(u_i|X_i) = 0$ holds.

4) OLS Assumptions

- Assumption 1: $E(u_i|X_i) = 0$.

$$(i) X_i \text{ and } u_i \text{ are independent and } E(u_i) = 0$$

$$\Downarrow (Yes) \quad \not\Uparrow \textbf{(No)}$$

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

$$\Downarrow (Yes) \quad \not\Uparrow \textbf{(No)}$$

$$(iii) cov(u_i, X_i) = 0$$

- Assumption 1 is a strong assumption. It can fail easily. In practice, it may be hard to justify.
- Fortunately, econometricians have developed many other methods for the case where Assumption 1 does not hold.

4) OLS Assumptions

- **Assumption 2**: (X_i ,Y_i) are i.i.d.

- This arises automatically if the entity (individual, district) is sampled by simple random sampling: the entity is selected then, for that entity, X and Y are observed (recorded).

- The main place we will encounter non-i.i.d. sampling is when data are recorded over time ("time series data") – this will introduce some extra complications.

4) OLS Assumptions

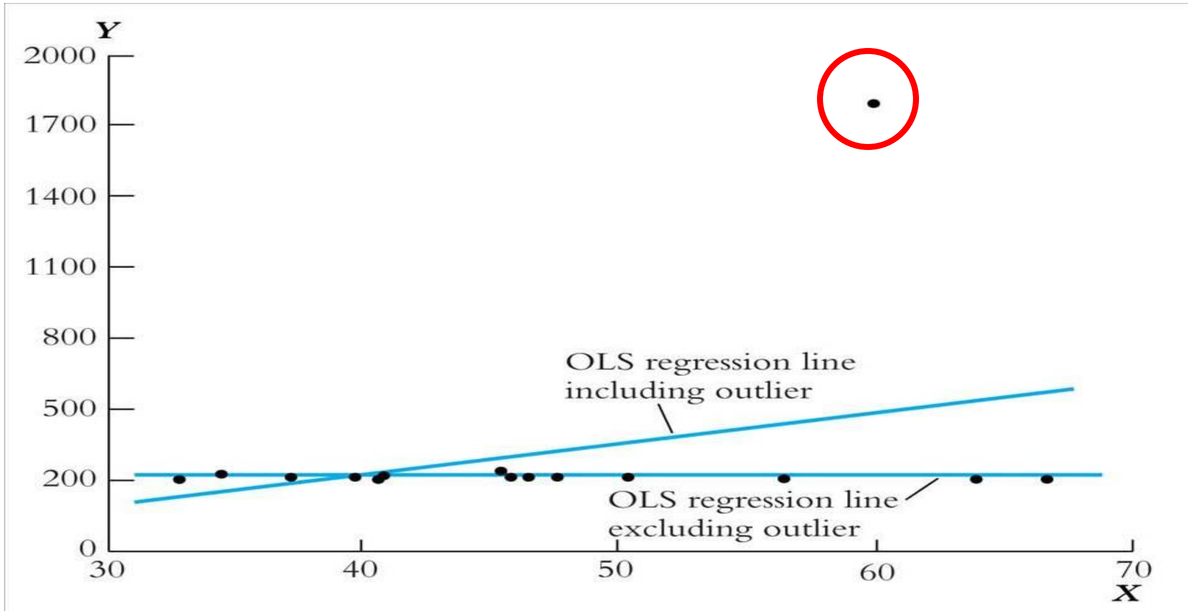
- **Assumption 3**: outliers are unlikely.
- Technically,

$$0 < E(X_i^4) < \infty \text{ and } 0 < E(Y_i^4) < \infty$$

- 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).
- However, the substance of this assumption is that a large outlier can strongly influence the results

4) OLS Assumptions

- Assumption 3: outliers are unlikely.



X_i	Y_i
10	3000
8	2400
6	2300

↓ *oops*

X_i	Y_i
10	30000
8	2400
6	2300

- Is the lone point an outlier in X or Y?
- In practice, outliers often are data glitches (coding/recording problems) – so check your data for outliers!

4) OLS Assumptions

Econ 3334

- Why do we need to make these assumptions?
- We need them to prove that OLS estimators are
  - Unbiased and Consistent
  - Asymptotically normally distributed (the distributions of OLS estimators are close to normal distribution)
  - Efficient (if we further assume homoskedasticity, $var(u_i|X_i)$=constant, i.e., conditional variance of u_i given X_i is constant)

THE GAUSS-MARKOV THEOREM FOR $\hat{\beta}_1$

If the three least squares assumptions in Key Concept 4.3 hold *and* if errors are homoskedastic, then the OLS estimator $\hat{\beta}_1$ is the **B**est (most efficient) **L**inear conditionally **U**nbiased **E**stimator (is **BLUE**).

5) Sampling Distribution of OLS estimators

- The data we use comes from a sample taken from a underlying population

- The data differ from sample to sample, and thus so does the OLS estimators $\hat{\beta}_0, \hat{\beta}_1$

- $\hat{\beta}_0, \hat{\beta}_1$ are random variables since they come from random samples (just like $\bar{Y}$ in Chapter 3)

- We need to understand their probability distribution, so we can make inferences (e.g., hypothesis testing, confidence interval)

5) Sampling Distribution of OLS estimators

- Like $\bar{Y}$, $\hat{\beta}_1$ has a sampling distribution.
- What is E($\hat{\beta}_1$)? (where is it centered?)

  If E($\hat{\beta}_1$) = β_1, then OLS is unbiased – a good thing!
- What is var($\hat{\beta}_1$)? (measure of sampling uncertainty)
- What is the distribution of in small samples?

  It can be very complicated in general
- What is the distribution of in large samples?

  It turns out to be relatively simple – in large samples, is normally distributed.

5) Sampling Distribution of OLS estimators

- Under the Key OLS Assumption 1: $E\left(\hat{\beta}_0\right) = \beta_0$ and $E\left(\hat{\beta}_1\right) = \beta_1$
- Given the OLS assumptions, the large sample distribution of $\hat{\beta}_0, \hat{\beta}_1$ are:

$$\hat{\beta}_1 \sim N\left(\beta_1, \sigma^2_{\hat{\beta}_1}\right),$$

where $\sigma^2_{\hat{\beta}_1} = \frac{1}{n}\frac{var[(X_i - \mu_X)u_i]}{[var(X_i)]^2}$

$$\hat{\beta}_0 \sim N\left(\beta_0, \sigma^2_{\hat{\beta}_0}\right),$$

where $\sigma^2_{\hat{\beta}_0} = \frac{1}{n}\frac{var(H_i u_i)}{[E(H_i^2)]^2}, \; H_i = 1 - \left(\frac{\mu_X}{E(X_i^2)}\right)X_i$

5) Sampling Distribution of OLS estimators

- We see that

$$var\left(\hat{\beta}_1\right) = \frac{1}{n}\frac{var[(X_i - \mu_X) \cdot u_i]}{[var(X_i)]^2}$$

$$var\left(\hat{\beta}_0\right) = \frac{1}{n}\frac{var(H_i u_i)}{[E(H_i^2)]^2}$$

- Variance of $\hat{\beta}_0$ and $\hat{\beta}_1$ are proportionate to n. Thus as n goes to infinite, the variance becomes smaller and smaller. This means that $\hat{\beta}_0$ and $\hat{\beta}_1$ become more and more concentrated around true β_0 and β_1.
- Thus $\hat{\beta}_0$ and $\hat{\beta}_1$ are consistent.