Measures of Fit

(Section 4.3)

A natural question is how well the regression line "fits" or explains the data. There are two regression statistics that provide complementary measures of the quality of fit:

- The ***regression 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)

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

The *regression* R^2 is the fraction of the sample variance of Y_i "explained" by the regression.

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

$\Rightarrow$ sample var (Y) = sample var($\hat{Y}_i$) + sample var($\hat{u}_i$) (*why?*)

$\Rightarrow$ total sum of squares = "explained" SS + "residual" SS

Definition of R^2:

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

- $R^2 = 0$ means $ESS = 0$
- $R^2 = 1$ means $ESS = TSS$
- $0 \le R^2 \le 1$
- For regression with a single X, R^2 = the square of the correlation coefficient between X and Y

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:

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

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

(the second equality holds because $\bar{\hat{u}} = \frac{1}{n}\sum_{i=1}^{n}\hat{u}_i = 0$).

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

The *SER*:

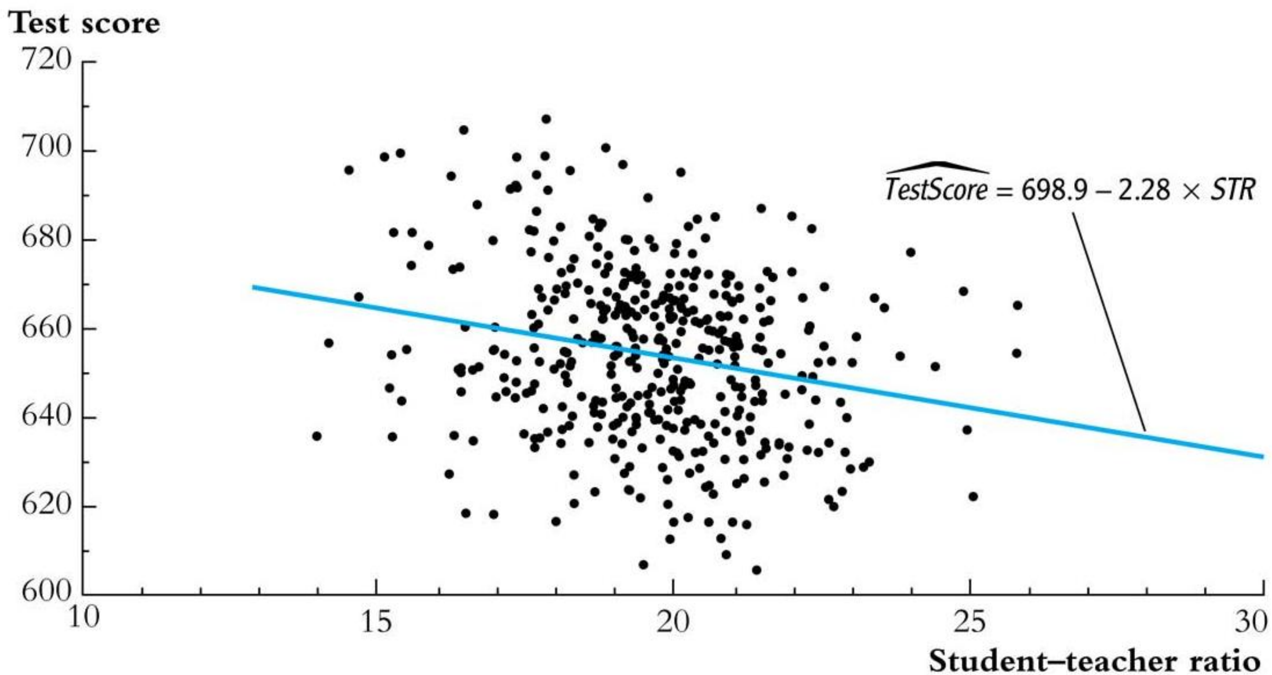
- has the units of *u*, which are the units of *Y*
- measures the average "size" of the OLS residual (the average "mistake" made by the OLS regression line)

Technical note: why divide by n–2 instead of n–1?

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

- Division by n–2 is a "degrees of freedom" correction – just like division by n–1 in s_Y^2, except that for the *SER*, two parameters have been estimated (β_0 and β_1, by $\hat{\beta}_0$ and $\hat{\beta}_1$), whereas in s_Y^2 only one has been estimated (μ_Y, by $\bar{Y}$).
- When n is large, it makes negligible difference whether n, n–1, or n–2 are used – although the conventional formula uses n–2 when there is a single regressor.

Example of the R^2 and the *SER*



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

STR explains only a small fraction of the variation in test scores. Does this make sense? Does this mean the STR is unimportant in a policy sense?