

Econ 312: Introduction to Econometrics

Simple Linear Regression

Sang-Yeob Lee

March 2, 2010

Measures of Fit

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 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.

$$Y_i = \hat{Y}_i + \hat{\epsilon}_i = \text{OLS Prediction} + \text{OLS residual}$$

- *Sample Var*(Y) = *sample Var*(\hat{Y}_i) + *sample Var*($\hat{\epsilon}_i$) (why?)
- $\sum_{i=1}^n (Y_i - \bar{Y})^2 = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^n \hat{\epsilon}_i^2$.
- total sum of square(TSS) = “explained” SS (ESS) + “residual” SS (RSS)

Ctd.

Definition of R^2

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

- $R^2 = 0$ means $ESS=0$
- $R^2 = 1$ means $ESS=TSS$
- $0 \leq R^2 \leq 1$
- 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 ϵ . The SER is (almost) the sample standard deviation of the OLS residuals:

$$\begin{aligned} SER &= \sqrt{\frac{1}{n-2} \sum_{i=1}^n (\hat{\epsilon}_i - \bar{\hat{\epsilon}})^2} \\ &= \sqrt{\frac{1}{n-2} \sum_{i=1}^n \hat{\epsilon}_i^2} \end{aligned}$$

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

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

The SER:

- has the units of ϵ , which are the units of Y .
- measures the average “size” of the OLS residual (the average “mistake” made by the OLS regression line.)
- The root mean square error (RMSE) is closely related to

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

This measures the same thing as the SER-minor difference is division by $1/n$ instead of $1/(n-2)$

Technical note: Why divide by $n - 2$ instead of $n - 1$?

- 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 $n - 2$ is used—although the conventional formula uses $n - 2$ when there is a single regressor.

Example of the R^2 and the SER

Estimated regression line: $\widehat{TestScore} = 689.9 - 2.28 \times STR$,
 $R^2 = 0.05$, $SER = 18.6$

- The $R^2 = 0.051$ means the regressor STR explains 5.1% of the variance of the dependent variable $TestScore$. STR explains a small fraction of the variation in test scores.
- The SER of 18.6 means that standard deviation of the regression residuals is 18.6. There is a large spread of the scatter plot around the regression line.
- The fact that the R^2 of this regression line is low (and the SER is large) does not, by itself, imply that this regression is either “bad” or “good”. What the low R^2 does tell us is that other importance factors influence test scores.

The Least Squares Assumptions

What, in a precise sense, are the properties of the OLS estimator? We would like it to be unbiased and to have a small variance. Does it? Under what conditions is it an unbiased estimator of the true population parameters?

To answer these questions, we need to make some assumptions about how Y and X are related each other, and about how they are corrected (the sampling scheme)

These assumptions are known as the least square assumptions.

The Classical Assumptions

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \quad i = 1, \dots, n$$

- 1 The conditional distribution of ϵ given X has mean zero, that is $E(\epsilon|X) = 0$
This implies that $\hat{\beta}_1$ is unbiased
- 2 $(X_i, Y_i), \quad i = 1, \dots, n$ 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 Y are rare.
 - Technically, X and Y have finite fourth moment (i.e. $E(X^4) < \infty$)
 - Outliers can result in meaningless values of $\hat{\beta}_1$.

$$E(\epsilon|X) = 0$$

- For any given value of X , the mean of ϵ is zero.
- Recall that if the conditional mean of one random variable given another is zero, then two random variables have zero covariance and thus are uncorrelated.
- Thus, the conditional mean assumption $E(\epsilon|X) = 0$ implies that X_i and ϵ_i are uncorrelated.

Ctd.

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 can be done by computer-using no information about individual.
- Because X is assigned randomly, all other individual characteristics-the things that make up ϵ -are independently distributed of X .
- Thus, in an ideal randomized controlled experiments, $E(\epsilon|X) = 0$.
- In actual experiments, or with observational data, we will need to think hard about whether $E(\epsilon|X)$ holds.

Least squares assumption 3. Large outliers are rare. Technical statement:

$$E(X^4) < \infty \text{ and } E(Y^4) < \infty$$

- A large outlier is an extreme value of X and Y .
- On a technical level, if X and Y are bounded then, they have finite fourth moments. (Standardized test score automatically satisfy this; STR, family income, etc. satisfy this too).
- However, the substance of this assumption is that a large outlier can strongly influence OLS estimator.
- In practices, outliers often are data glitches (coding/recording problems)-so check your data for outliers !. The easiest way is to produce a scatterplot.

The sampling distribution of the OLS estimator

The OLS estimator is computed from a sample of data; a different sample gives a different values of $\hat{\beta}_1$. This is the source of the “sampling uncertainty” of $\hat{\beta}_1$. We want to:

- quantify the sampling uncertainty associated with $\hat{\beta}_1$.
- use $\hat{\beta}_1$ to test hypotheses such as $\beta_1 = 0$
- construct a confidence interval for β_1 .
- All these require figuring out the sampling distribution of the OLS estimator. Two steps to get there...
 - Probability framework for linear regression
 - Distribution of the OLS estimator

Probability Framework for Linear Regression

The probability framework for linear regression is summarized by the three least squares assumptions.

- Population: The group of interest (ex: all possible school district)
- Random variables: Y, X
- Joint distribution of (Y, X)
 - The population regression function is linear.
 - $E(\epsilon|X)$ X and ϵ are uncorrelated.
 - X, Y has finite fourth moments.
- Data Collection by simple random sampling:
 $\{(X_i, Y_i)\}, \quad i = 1, \dots, n$ are i.i.d.

The sampling distribution of $\hat{\beta}_1$

Like the sample mean \bar{Y} , $\hat{\beta}_1$ has a sampling distribution.

- What is $E(\hat{\beta}_1)$? (Where is it centered?)
 - If $E(\hat{\beta}_1) = \beta_0$, then OLS is unbiased—a good thing!
- What is $Var(\hat{\beta}_1)$? (measure of sampling uncertainty)
- What is the distribution of $\hat{\beta}_1$ in small samples?
 - It can be very complicated in general.
- What is the distribution of $\hat{\beta}_1$ in large sample?
 - It turns out to be relatively simple—in large samples, $\hat{\beta}_1$ is normally distributed.

The mean and variance of the sampling distribution of $\hat{\beta}_1$

Some preliminary algebra:

$$\begin{aligned}Y_i &= \beta_0 + \beta_1 X_i + \epsilon_i \\ \bar{Y} &= \beta_0 + \beta_1 \bar{X} + \bar{\epsilon} \\ \text{so, } Y_i - \bar{Y} &= \beta_1 (X_i - \bar{X}) + (\epsilon_i - \bar{\epsilon})\end{aligned}$$

Thus,

$$\begin{aligned}\hat{\beta}_1 &= \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2} \\ &= \frac{\sum_{i=1}^n (X_i - \bar{X})[\beta_1 (X_i - \bar{X}) + (\epsilon_i - \bar{\epsilon})]}{\sum_{i=1}^n (X_i - \bar{X})^2}\end{aligned}$$

$$\begin{aligned}\hat{\beta}_1 &= \frac{\sum_{i=1}^n (X_i - \bar{X})[\beta_1(X_i - \bar{X}) + (\epsilon_i - \bar{\epsilon})]}{\sum_{i=1}^n (X_i - \bar{X})^2} \\ &= \beta_1 \frac{\sum_{i=1}^n (X_i - \bar{X})(X_i - \bar{X})}{\sum_{i=1}^n (X_i - \bar{X})^2} + \frac{\sum_{i=1}^n (X_i - \bar{X})(\epsilon_i - \bar{\epsilon})}{\sum_{i=1}^n (X_i - \bar{X})^2}\end{aligned}$$

so, $\hat{\beta}_1 - \beta_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(\epsilon_i - \bar{\epsilon})}{\sum_{i=1}^n (X_i - \bar{X})^2}$ Now,

$$\begin{aligned}\sum_{i=1}^n (X_i - \bar{X})(\epsilon_i - \bar{\epsilon}) &= \sum_{i=1}^n (X_i - \bar{X})\epsilon_i - \left[\sum_{i=1}^n (X_i - \bar{X}) \right] \bar{\epsilon} \\ &= \sum_{i=1}^n (X_i - \bar{X})\epsilon_i - \left[\left(\sum_{i=1}^n X_i \right) - n\bar{X} \right] \bar{\epsilon} \\ &= \sum_{i=1}^n (X_i - \bar{X})\epsilon_i\end{aligned}$$

Substitute $\sum_{i=1}^n (X_i - \bar{X})(\epsilon_i - \bar{\epsilon}) = \sum_{i=1}^n (X_i - \bar{X})\epsilon_i$ into the expression for $\hat{\beta}_1 - 1 - \beta_1$:

$$\hat{\beta}_1 - \beta_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(\epsilon_i - \bar{\epsilon})}{\sum_{i=1}^n (X_i - \bar{X})^2}$$

so,

$$\hat{\beta}_1 - \beta_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})\epsilon_i}{\sum_{i=1}^n (X_i - \bar{X})^2}$$

Now we can calculate $E(\hat{\beta}_1)$ and $Var(\hat{\beta}_1)$

$$\begin{aligned} E(\hat{\beta}_1) - \beta_1 &= E\left[\frac{\sum_{i=1}^n (X_i - \bar{X})\epsilon_i}{\sum_{i=1}^n (X_i - \bar{X})^2}\right] \\ &= E\left\{E\left[\frac{\sum_{i=1}^n (X_i - \bar{X})\epsilon_i}{\sum_{i=1}^n (X_i - \bar{X})^2} \middle| X_1, \dots, X_n\right]\right\} \\ &= 0 \end{aligned}$$

because $E(\epsilon_i | X_i = x) = 0$ by the LSA #1.

- Thus, LSA # implies that $E(\hat{\beta}_1) = 0$
- That is, $\hat{\beta}_1$ is an unbiased estimator of β_1 .

Next calculate $\text{Var}(\hat{\beta}_1)$

write

$$\hat{\beta}_1 - \beta_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})\epsilon_i}{\sum_{i=1}^n (X_i - \bar{X})^2} = \frac{\frac{1}{n} \sum_{i=1}^n v_i}{\frac{n-1}{n} s_X^2}$$

where $v_i = (X_i - \bar{X})\epsilon_i$. If n is large, $s_X^2 \approx \sigma_X^2$, and $\frac{n-1}{n} \approx 1$, so,

$$\hat{\beta}_1 - \beta_1 \approx \frac{\frac{1}{n} \sum_{i=1}^n v_i}{\sigma_X^2}$$

where $v_i = (X_i - \bar{X})\epsilon_i$. Thus,

$$\text{Var}(\hat{\beta}_1 - \beta_1) = \text{Var}(\hat{\beta}_1) = \frac{\text{Var}(v)/n}{(\sigma_X^2)^2}$$

Thus,

$$\text{Var}(\hat{\beta}_1 - \beta_1) = \text{Var}(\hat{\beta}_1) = \frac{1}{n} \frac{\text{Var}[(X_i - \mu_X)\epsilon_i]}{(\sigma_X^2)^2}$$

What is the sampling distribution of $\hat{\beta}_1$?

The exact sampling distribution is complicated-it depends on the population distribution of (Y, X) -but when n is large we get some simple and good approximations:

- ① Because $\text{Var}(\hat{\beta}_1) \propto 1/n$ and $E(\hat{\beta}_1) = \beta_1$, $\hat{\beta}_1 \rightarrow^p \beta_1$
- ② When n is large, the sampling distribution of $\hat{\beta}_1$ is well approximated by a normal distribution (CLT)

Recall the CLT: suppose $v_i, i = 1, \dots, n$ is i.i.d, with $E(v) = 0$ and $\text{Var}(v_i) = \sigma_v^2$. Then, when n is large, $\frac{1}{n} \sum_{i=1}^n v_i$ is approximately distributed $N(0, \sigma_v/n)$

Large- n approximation to the distribution of $\hat{\beta}_1$

$$\hat{\beta}_1 - \beta_1 \approx \frac{\frac{1}{n} \sum_{i=1}^n v_i}{\sigma_X^2}$$

- When n is large, $v_i = (X_i - \bar{X})\epsilon_i \approx (X_i - \mu_X)\epsilon_i$, which is i.i.d and $\text{Var}(v_i) < \infty$. So, by the CLT,
- $\frac{1}{n} \sum_{i=1}^n v_i$ is approximately distributed $N(0, \sigma_v^2/n)$
- Thus, for n large, $\hat{\beta}_1$ is approximately distributed.

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma_v^2}{n\sigma_X^4}\right)$$

, where $v_i = (X_i - \bar{X})\epsilon_i$.

The larger the variance of X , the smaller the variance of $\hat{\beta}_1$

The math:

$$\text{Var}(\hat{\beta}_1 - \beta_1) = \text{Var}(\hat{\beta}_1) = \frac{1}{n} \frac{\text{Var}[(X_i - \mu_X)\epsilon_i]}{(\sigma_X^2)^2}$$

where $\sigma_X^2 = \text{Var}(X_i)$. The variance of X appears in its square in the denominator-so increasing the spread of X decreases the variance of β_1 .

The intuition:

If there is more variation in X , then there is more information in the data that you can use to fit the regression line.

Summary of the sampling distribution of $\hat{\beta}_1$

If the three Least Squares Assumptions hold, then

- The exact sampling distribution of $\hat{\beta}_1$ has:
 - $E(\hat{\beta}_1) = \beta_1$ (unbiased)
 - $Var(\hat{\beta}_1) = \frac{1}{n} \frac{Var[(X_i - \mu_X)\epsilon_i]}{(\sigma_X^2)^2} \propto \frac{1}{n}$.
- Other than its mean and variance, the exact distribution of $\hat{\beta}_1$ is complicated and depends on the distribution of (X, ϵ) .
- $\hat{\beta}_1 \rightarrow^p \beta_1$ (That is, $\hat{\beta}_1$ is consistent.)
- When n is large, $\frac{\hat{\beta}_1 - E(\hat{\beta}_1)}{\sqrt{Var(\hat{\beta}_1)}} \sim N(0, 1)$ (CLT).
- This parallels the sampling distribution of \bar{Y}

We are now ready to turn to hypothesis tests and confidence interval...