The Simple Linear Regression Model

ECONOMETRICS (ECON 360)

BEN VAN KAMMEN, PHD

Outline

Definition.

Deriving the Estimates.

Properties of the Estimates.

Units of Measurement and Functional Form.

Expected Values and Variances of the Estimators.

Definition of Simple Linear Regression

Correlation: measures the “strength” of a linear relationship between two variables.

Regression: measures the way the expectation of one (“dependent”) variable changes when another (“independent”) variable changes.

Formally, estimate what is:

$$\frac{dE(y)}{dx}$$, where "d" represents "change in".

- A linear trend line on the scatterplot.
- Regression estimates the slope.
- “Fit” a straight line as closely as possible to the data points.

Regression parameters

Trend line form:

$$y = [Intercept] + [slope] * x.$$

Define parameters for the intercept and slope;

- use calculus to estimate them.
- β_0 as the intercept and β_1 as the slope. Then,

$$y = \beta_0 + \beta_1 x.$$

Residuals

$$E(y_i|x_i) = \beta_0 + \beta_1 x_i$$

gives you the expectation of y conditional on x.

Unless all the data points lie on a perfect line, there is a residual distance between the trend line and each data point, $[y_i - E(y_i|x_i)]$.

- Sometimes called the "error term".
- Denoted u_i:

$$u_i \equiv [y_i - E(y_i|x_i)].$$

Stochastic and non-stochastic

y_i can be decomposed into two parts:

1) the part that can be explained by y's relationship with x, and

2) the residual that cannot be explained by the relationship with x.

$$u_i \equiv [y_i - E(y_i|x_i)] \rightarrow y_i = E(y_i|x_i) + u_i \Leftrightarrow y_i = \beta_0 + \beta_1 x_i + u_i$$

Parameters and statistics

The model on the previous slide contains population parameters—as if their values were known.

In reality they are not known and must be inferred from a sample.

- Just like population mean, proportion and variance.

The estimators we use to infer these values are:

$$\hat{\beta}_0 \text{ , which estimates } \beta_0, \text{and } \hat{\beta}_1, \text{which estimates } \beta_1 \text{ .}$$

The residual estimates for the sample are $\hat{u}_i$.

Deriving Ordinary Least Squares (OLS) estimates

Estimating the parameters (slope and intercept) relies on calculus,

- as does every problem in economics in which you try to optimize something,
- e.g., utility maximization or cost minimization.
- In this application we minimize the sum of the squares of the residuals and therefore the distance between the line and the data points.

Sum of squared residuals

Re-arrange the function relating y to x:

$$\hat{u}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \rightarrow \hat{u}_i{}^2 = \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2.$$

$$\text{Sum of Squared Residuals } (SSR) = \sum_{i=1}^{n} \hat{u}_i{}^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

Sum of squared residuals (continued)

$$SSR = \sum_{i=1}^{n} (y_i^2 + \hat{\beta}_0^2 + \hat{\beta}_1^2 x_i^2 - 2\hat{\beta}_0 y_i - 2\hat{\beta}_1 x_i y_i + 2\hat{\beta}_0 \hat{\beta}_1 x_i)$$

The previous line squares the term in the sum. The next one expands the sum.

- Since the beta hat terms are not indexed with *i*, they can be pulled through the sums.

$$SSR = \sum_{i=1}^{n} y_i^2 + n\hat{\beta}_0^2 + \hat{\beta}_1^2 \sum_{i=1}^{n} x_i^2 - 2\hat{\beta}_0 \sum_{i=1}^{n} y_i - 2\hat{\beta}_1 \sum_{i=1}^{n} x_i y_i + 2\hat{\beta}_0 \hat{\beta}_1 \sum_{i=1}^{n} x_i$$

Minimizing SSR

By differentiating the above line with respect to the two statistics, $\hat{\beta}_0$ and $\hat{\beta}_1$.

The first order conditions for this minimum are:

$$\frac{\partial SSR}{\partial \hat{\beta}_0} = 0 \text{ and } \frac{\partial SSR}{\partial \hat{\beta}_1} = 0$$

Solving them simultaneously gives you the estimates.

$$(1)\ \frac{\partial SSR}{\partial \hat{\beta}_0} = 2n\hat{\beta}_0 - 2\sum_{i=1}^{n} y_i + 2\hat{\beta}_1 \sum_{i=1}^{n} x_i$$

$$(2)\ \frac{\partial SSR}{\partial \hat{\beta}_1} = 2\hat{\beta}_1 \sum_{i=1}^{n} {x_i}^2 - 2\sum_{i=1}^{n} x_i y_i + 2\hat{\beta}_0 \sum_{i=1}^{n} x_i$$

Minimizing SSR (continued)

Setting (1) equal to zero and solving for $\hat{\beta}_0$:

$$2n\hat{\beta}_0 - 2\sum_{i=1}^{n} y_i + 2\hat{\beta}_1 \sum_{i=1}^{n} x_i = 0 \rightarrow \hat{\beta}_0 = \frac{1}{n}\left[\sum_{i=1}^{n} y_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i\right] = \bar{y} - \hat{\beta}_1 \bar{x}$$

Substituting this into condition (2) gives you:

$$2\hat{\beta}_1 \sum_{i=1}^{n} x_i^2 - 2\sum_{i=1}^{n} x_i y_i + 2(\bar{y} - \hat{\beta}_1 \bar{x}) \sum_{i=1}^{n} x_i$$

Minimizing SSR (continued)

Setting the above expression equal to zero and solving for $\hat{\beta}_1$ gives you:

$$2\hat{\beta}_1 \sum_{i=1}^{n} {x_i}^2 - 2\sum_{i=1}^{n} x_i y_i + 2\left(\bar{y} - \hat{\beta}_1 \bar{x}\right) \sum_{i=1}^{n} x_i = 0 \rightarrow \hat{\beta}_1 \left(\sum_{i=1}^{n} {x_i}^2 - \bar{x} \sum_{i=1}^{n} x_i \right)$$

$$= \left(\sum_{i=1}^{n} x_i y_i - \bar{y} \sum_{i=1}^{n} x_i \right), \text{and}$$

$$\hat{\beta}_1 = \frac{\left(\sum_{i=1}^{n} x_i y_i - \bar{y} \sum_{i=1}^{n} x_i\right)}{\left(\sum_{i=1}^{n} x_i^2 - \bar{x} \sum_{i=1}^{n} x_i\right)}.$$

Simplifying the OLS estimates

We can further simplify this using the definition of $\bar{x}$:

$$\bar{x} \equiv \frac{1}{n}\sum_{i=1}^{n} x_i \Leftrightarrow n\bar{x} \equiv \sum_{i=1}^{n} x_i$$

$$\hat{\beta}_1 = \frac{\left(\sum_{i=1}^{n} x_i y_i - \bar{y} n \bar{x}\right)}{\left(\sum_{i=1}^{n} {x_i}^2 - \bar{x} n \bar{x}\right)} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Simplifying the OLS estimates (continued)

It can be shown that the numerator in this expression is equal to:

$$\left(\sum_{i=1}^{n} x_i y_i - \bar{y} n \bar{x}\right) = s_{xy}(n-1)$$

And the denominator equals:

$$\left(\sum_{i=1}^{n} {x_i}^2 - \bar{x} n \bar{x}\right) = s_x^2(n-1)$$

So the slope of the regression line is:

$$\hat{\beta}_1 = \frac{s_{xy}(n-1)}{s_x^2(n-1)} \text{ or } \hat{\beta}_1 = \frac{s_{xy}}{s_x^2} = \frac{Covariance}{Variance\ of\ x}.$$

And the intercept is:

$$\hat{\beta}_0 = \bar{y} - \frac{s_{xy}}{s_x^2}\bar{x}$$

Correlation and regression (concluded)

Since the regression coefficient, $\hat{\beta}_1$, and the correlation coefficient are both related to covariance, there is a relationship between regression and correlation, too. Specifically,

$$\hat{\beta}_1 = \frac{s_{xy}}{s_x^2} = \frac{r_{xy} s_x s_y}{s_x^2}, \text{where } r_{xy} \text{ is the correlation coefficient.}$$

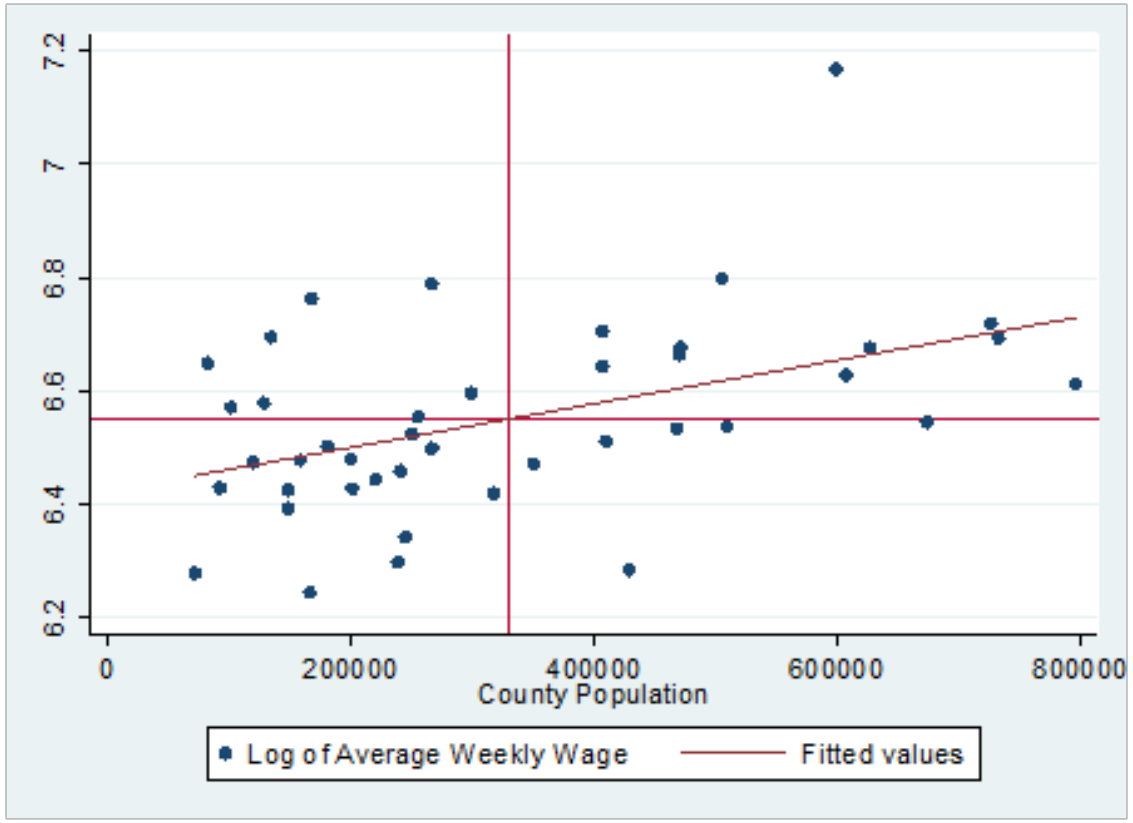
Since standard deviation is the square root of the variance, we can simplify this to:

$$\hat{\beta}_1 = \frac{r_{xy} s_y}{s_x}.$$

Example from the review

In the case of the population, wage relationship from earlier, the regression slope is 0.0385.

Suggests that on average, an extra 100,000 population increases the weekly wage by 3.85 log points, or roughly 3.85%.



OLS derivation concluded

The upward-sloping line is the linear regression estimate.

Note that the line goes through the point $(\bar{x}, \bar{y})$.

This suggests that we can fit the same line on a "de-meaned" data set that will have an intercept of zero.

Putting together both estimates, we can specify the expected value of y conditional on x.

$$E(y_i|x_i) = \bar{y} - \frac{S_{xy}}{S_x^2}\bar{x} + \frac{S_{xy}}{S_x^2}x_i \Leftrightarrow E(y_i|x_i) - \bar{y} = \frac{S_{xy}}{S_x^2}(x_i - \bar{x})$$

Properties of OLS on any sample of data

The observed value of y can be broken into an "explained" part and an "unexplained" part.

- The textbook calls the conditional expectation, "$\hat{y}_i$".
- This is also sometimes called a "fitted value" of y. The estimated residual, then is:

$$\hat{u}_i = [y_i - \hat{y}_i] = (y_i - \bar{y}) + (\bar{y} - \hat{y}_i) \Leftrightarrow (y_i - \bar{y}) = \hat{u}_i + (\hat{y}_i - \bar{y})$$

$$(y_i - \bar{y})^2 = \hat{u}_i{}^2 + (\hat{y}_i - \bar{y})^2 + 2\hat{u}_i(\hat{y}_i - \bar{y})$$

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}\hat{u}_i{}^2 + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + 2\sum_{i=1}^{n}\hat{u}_i(\hat{y}_i - \bar{y})\,.$$

The last term sums to zero. This digression shows this.

Variance decomposition

So the sum of squared deviations from the mean ("SST") is the sum of the sum of squared residuals ("SSR") and the sum of squares explained by regression ("SSE").

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}\hat{u}_i{}^2 + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 \text{ or } SST = SSR + SSE$$

Coefficient of determination

The strength of a regression model can be measured by the proportion of y's total variation that can be explained by x. This is called the coefficient of determination (r^2).

$$r^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{\left(\frac{s_{xy}}{s_x^2}\right)^2 (n-1)s_x^2}{(n-1)s_y^2} = \frac{{s_{xy}}^2}{s_x^2 s_y^2}$$

The square root of the coefficient of determination is:

$$\sqrt{r^2} = \sqrt{\frac{{s_{xy}}^2}{s_x^2 s_y^2}} = \frac{s_{xy}}{s_x s_y} = r_{xy} \text{ the correlation coefficient!}$$

OLS properties concluded

Make sure that you take the sign (+/-) from the regression coefficient for this square root, or else all correlations will seem to be positive.

For the regression of wage on population, the coefficient of determination is 0.2101—which is the square of the correlation coefficient, 0.4583.

Units of measurement and functional form

A fundamental skill that all empiricists must have is being able to interpret the meaning of their results.

OLS estimates of the slope and intercept parameters have specific interpretations:

- namely that the slope coefficient estimates the effect (on y) of increasing x by one unit, and
- the intercept estimates the expected value of y when x equals zero.

$$\Delta y = \beta_1 \Delta x;\ \Delta x = 1 \rightarrow \Delta y = \beta_1 \text{ and } y(|x = 0) = \beta_0 + 0.$$

Units of measurement

Consider a regression model estimating how long a labor dispute ("strike") will last.

- This could be measured in days, weeks, hours, nanoseconds, etc.

Say that your model is using the average labor market experience of the union members to explain how long their strike lasts (x=average experience in years).

Initially you measure duration of the strike in hours.

- Accordingly β_1 measures the additional number of hours the strike is expected to last when the union involved has 1 extra year (on average) of experience.

$$\Delta x = 1 \text{ year} \rightarrow \Delta y = \beta_1 \text{ hours.}$$

Units of measurement (continued)

Now imagine you measure duration of the strikes in *days* instead.

$$24 \text{ hours} = 1 \text{ day; if } y \equiv \text{hours and } y' \equiv \text{days, then } y' = \frac{\text{y}}{24}.$$

So,

$$y = 24y'.$$

Plug this into the regression model that uses hours to see how to interpret the new effect of an extra year of experience.

$$y' = \frac{\beta_0}{24} + \frac{\beta_1}{24}x + u;\ \Delta x = 1 \text{ year} \rightarrow \Delta y' = \frac{\beta_1}{24} \text{ days.}$$

Units of measurement (concluded)

The effect is now precisely (and unsurprisingly) 1/24 as large as before!

To make this concrete, if you regress hours of duration on average experience and the coefficient estimate is ($\hat{\beta}_1 = 72$ hours), a regression of *days* of duration on average experience should yield an estimate of exactly ($\hat{\beta}_1 = 3$ days).

The estimated effect does not actually change when you change units of measurement—because you're just scaling one of the variables by a constant!

- Only the interpretation changes.
- The lesson is merely (but crucially) to be cognizant of the units of measurement whenever interpreting regression results.

Other variable transformations

This principle can be generalized to *transformations* that do not involve multiplying by a constant.

For example the *natural logarithm*,

- frequently done in earnings regressions in labor economics ($y \equiv \ln(\$\, earnings)$).
- For simplicity again imagine you are explaining earnings using labor market *experience* so that:

$$y = \beta_0 + \beta_1 experience + u;\ \Delta x = 1\ \text{year} \rightarrow \Delta y = \beta_1.$$

Logarithms and percent change

But what does Δy *mean*?

It does not mean the additional $ of earnings, because y isn't measured in dollars.

Instead,

$$\Delta y = \ln(y_1) - \ln(y_0)\,,$$

which is an approximation of the $\%\Delta$ in $earnings$.

Logs and percent change (continued)

Recall that,

$$\%\Delta earnings \equiv 100 * \frac{earnings_1 - earnings_0}{earnings_0} = 100 * \left(\frac{earnings_1}{earnings_0} - 1\right), \text{so}$$

$$1 + \frac{\%\Delta earnings}{100} = \frac{earnings_1}{earnings_0}. \text{ Taking the log of both sides gives you,}$$

$$\ln\left(1 + \frac{\%\Delta earnings}{100}\right) = \ln(earnings_1) - \ln(earnings_0) = \Delta y = \beta_1.$$

Logs and percent change (concluded)

Taking the "anti-log" of β_1 shows you how to interpret the estimate as a $\%\Delta$.

$$\ln\left(e^{\beta_1}\right) = \beta_1 \rightarrow e^{\beta_1} = 1 + \frac{\%\Delta earnings}{100} \Leftrightarrow 100 * \left(e^{\beta_1} - 1\right) = \%\Delta earnings.$$

There will be numerous opportunities to practice this in homework exercises, but here is one concrete example. If your estimate of β_1 is 0.05, you have estimated a % increase in earnings (resulting from an extra year of experience) of:

$$\%\Delta earnings = 100 * (e^{0.05} - 1) = 5.13\%.$$

Log-level models

Estimating the effect of level of experience on the log of wage:

- (called a “log-level” regression),
- estimates a constant rate of return on experience, rather than a constant increase.
- They have proven much more appropriate in settings like earnings (among many others) because employees usually get % pay raises instead of fixed dollar amounts per year.

One more way of explaining the $\%\Delta$ interpretation of β_1 in a “log-level” regression uses the chain rule of calculus.

Log-log models

The last example of regression using transformations we will examine here is the "log-log" model, in which x and y are expressed in logarithms.

Building on the last result (using the chain rule), if the model we estimate is:

$$\ln(y) = \beta_0 + \beta_1 \ln(x) + u,$$

Then,

$$\rightarrow \frac{dln(y)}{dln(x)} = \beta_1 = \frac{dln(y)}{dy}\frac{dy}{dx}\frac{dx}{dln(x)} = \frac{1}{y}\frac{dy}{dx}x.$$

Log-log models: constant elasticity

The last equality can be re-arranged:

$$\beta_1 = \frac{dy}{y}\frac{x}{dx} = \frac{\%\Delta y}{\%\Delta x},$$

which any good economics student should recognize as anelasticity.

So a regression with both variables transformed estimates a constant elasticity relationship between x and y: "if x increases by 1%, y changes by $\beta_1\%$."

Expected value and variance of OLS estimators: under SLR assumptions 1-4

Any statistic calculated from a sample has an expected value and a standard error.

- OLS estimation relies on some assumptions to derive its expectation and standard error.

Explicitly one of the ones OLS makes is that the model is linear in its parameters:

$$(\text{SLR. } 1)\ y = \beta_0 + \beta_1 x + u, \text{in the population.}$$

This analysis has also assumed that we have estimated the population with a random sample, i.e., one that is representative of the population (if it were drawn many times).

- This is assumption (SLR.2) in the text.

OLS assumptions (continued)

We have assumed that the estimator is defined by assuming that there is variation in x (SLR.3). If this was not the case, the denominator of β_1 would be zero.

Since the linear model includes an intercept, we have essentially de-meaned the error term, u, and made its expected value zero.

- This assumption has shown up in the derivation of R^2 already. We can make it by imagining that any non-zero mean in the error could simply be added to the intercept term.

OLS assumptions (continued)

For several reasons it is necessary to assume mean independence between the error term and x:

$\hat{\beta}_1$ is an unbiased estimator (as we will show) if this assumption holds and the model requires it for $\hat{y}$ to accurately estimate the expected value of y conditional on x.

In order for this to hold, the model,

$$y = \beta_0 + \beta_1 x + u, \text{must satisfy}$$
$$E(y|x) = \beta_0 + \beta_1 x + 0.$$

OLS assumptions (concluded)

Thus $E(u|x)$ must equal zero (Assumption SLR.4).

- Combined with the zero mean assumption for the error term (unconditionally), you have the full zero conditional mean assumption:

$$E(u|x) = E(u) = 0.$$

Now to verify the claim preceding SLR.4 (that under these assumptions $\hat{\beta}_1$ is unbiased).

To be unbiased, the expected value of beta hat should equal the population parameter, β_1:

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

Expected value of the OLS estimator

The estimator is calculated as:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

Substitute the regression equation in for y_i.

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(\beta_0 + \beta_1 x_i + u_i)}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Expected value of the OLS estimator (continued)

Dwelling on the numerator for a moment, it can be multiplied out as follows:

$$\sum_{i=1}^{n}(x_i - \bar{x})(\beta_0 + \beta_1 x_i + u_i) = \beta_0 \sum_{i=1}^{n}(x_i - \bar{x}) + \beta_1 \sum_{i=1}^{n}(x_i - \bar{x})x_i + \sum_{i=1}^{n}(x_i - \bar{x})u_i \,.$$

Simplify this.

- The first term is zero because the sum of deviations from the sample mean is zero.
- The x terms in the second term are actually the same as the denominator, too:

$$\sum_{i=1}^{n}(x_i - \bar{x})x_i = \sum_{i=1}^{n} x_i^2 - \bar{x}n\bar{x} = \sum_{i=1}^{n}(x_i - \bar{x})^2 \,.$$

Expected value of the OLS estimator (continued)

And the last term simplifies to:

$$\sum_{i=1}^{n}(x_i - \bar{x})u_i = \sum_{i=1}^{n} x_i u_i - \bar{x} * \sum_{i=1}^{n} u_i = \sum_{i=1}^{n} x_i u_i - 0.$$

So now we have:

$$\hat{\beta}_1 = \frac{\beta_1 \sum_{i=1}^{n}(x_i - \bar{x})^2 + \sum_{i=1}^{n} x_i u_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \beta_1 + \frac{\sum_{i=1}^{n} x_i u_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

Expectation and unbiasedness

To show un-biasedness, take the expectation.

$$E(\hat{\beta}_1) = \beta_1 + E\left[\frac{\sum_{i=1}^{n} x_i u_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right].$$

The second term of the expression above will be zero if the numerator is zero, which it will be if:

$$E(x_i u_i) = E(x) * E(u|x) = 0.$$

Unbiasedness of OLS

If assumption SLR.4 holds, this is true, the second term is zero, and the expectation of $\hat{\beta}_1$ is β_1 (unbiased).

If any of the 4 assumptions is violated, OLS estimators will be biased.

In this introductory lecture, it is beyond our scope to discuss the direction of bias and what can be done to eliminate it.

- We postpone most of that discussion for the remainder of the course, but for now suffice it to say that the most tenuous assumption in econometric applications is SLR.4.

Variance of the OLS estimator

Measures how far from its mean the estimator is likely to be in any given sample.

Even if a statistic is unbiased, a finite sample estimate could still be “unlucky” and vary around the true value.

Here we show that the standard error of $\hat{\beta}_1$ is:

$$se(\hat{\beta}_1) \equiv Var(\hat{\beta}_1)^{\frac{1}{2}} = \sqrt{\frac{\sum_{i=1}^{n} \hat{u}_i{}^2}{(n-2)\sum_{i=1}^{n}(x_i - \bar{x})^2}} = \frac{SSR}{(n-2)} * \left[\sum_{i=1}^{n}(x_i - \bar{x})^2\right]^{-1}$$

Homoskedasticity

The variance of beta 1 is further equal to:

$$Var(\hat{\beta}_1) = \frac{SSR}{(n-2)} * \left[\sum_{i=1}^{n}(x_i - \bar{x})^2\right]^{-1} = \hat{\sigma}^2 \left[\sum_{i=1}^{n}(x_i - \bar{x})^2\right]^{-1},$$

where $\hat{\sigma}^2$ is an estimate of the variance of the error (u) under the assumption of homoskedasticity, i.e.,

$$Var(u|x) = Var(u) = \sigma^2.$$

Variance of the OLS estimator (continued)

Variance is the expected squared deviation from a random variable's mean (expected value).

$$Var(x) \equiv E\big(x - E(x)\big)^2 \rightarrow Var\big(\hat{\beta}_1\big) = E\left[\beta_1 + \frac{\sum_{i=1}^{n}(x_i - \bar{x})u_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2} - \beta_1\right]^2.$$

Most introductory econometrics, including this one, an additional simplification is used: treating x as a non-random variable. The consequence of treating x as a fixed regressor is that the variability of the estimator will come only from the error term.

$$Var\big(\hat{\beta}_1\big) = E\left[\frac{\sum_{i=1}^{n}(x_i - \bar{x})u_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]^2 = E\left[\frac{\sum_{i=1}^{n} d_i u_i}{SST_x}\right]^2 ; d_i \equiv x_i - \bar{x}.$$

Variance of the OLS estimator (continued)

The constant (total sum of squares of x) can now be pulled through the expectation, and attention can be focused on the sum in the numerator.

$$Var(\hat{\beta}_1) = {SST_x}^{-2} * E\left[\sum_{i=1}^{n} d_i u_i\right]^2$$

Squaring the sum of n terms means summing all the possible cross-products, i.e.,

$$(d_1u_1 + \ldots d_nu_n) * (d_1u_1 + \ldots d_nu_n) = d_1^2u_1^2 + d_1u_1d_2u_2 + \ldots d_n^2u_n^2.$$

Variance of the OLS estimator (continued)

There will be n "own" products (each element times itself) and $\frac{n(n-1)}{2}$ unique "cross" products, i.e., $d_i u_i * d_{-i} u_{-i}$.

There are 2 of each of the latter group. So,

$$Var(\hat{\beta}_1) = {SST_x}^{-2} * \left[E \sum_{i=1}^{n} d_i^2 u_i^2 + 2 * E(\text{sum of cross products}) \right].$$

Variance of the OLS estimator (continued)

Conveniently the sum of the cross products has an expectation of zero because the draws of the sample are independent, i.e., the correlation between observations 1 and 21 (or any pair of randomly chosen observations) is zero.

So now we are just down to the "own" products:

$$Var(\hat{\beta}_1) = {SST_x}^{-2} * E\sum_{i=1}^{n} d_i^2 u_i^2 = {SST_x}^{-2} * \sum_{i=1}^{n} d_i^2 E(u_i^2).$$

Variance of the OLS estimator (continued)

Since the expectation of u is zero, the expectation of u^2 is the variance of u.

Under homoskedasticity, this is the constant, σ^2, which can be pulled through the sum:

$$Var(\hat{\beta}_1) = {SST_x}^{-2} * \sum_{i=1}^{n} d_i^2 \sigma^2 = {SST_x}^{-2} * \sigma^2 \sum_{i=1}^{n} d_i^2 .$$

Lastly note that the sum of d_i^2 equals SST_x, so you can simplify as follows.

$$Var(\hat{\beta}_1) = {SST_x}^{-2} * \sigma^2 * SST_x = \frac{\sigma^2}{SST_x} = \frac{\sigma^2}{(n-1)(\text{variance of } x)}.$$

Variance of the OLS estimator (concluded)

The last equality serves to remind us of what happens to estimates as the sample size grows. "n" appears only in the denominator, so as it gets large, the variance of beta 1 gets small and the estimator gets more precise.

In order to use this measure practically, an estimate of the error variance is needed, though, which we have in the form of:

$$\hat{\sigma}^2 \equiv \frac{1}{n-2}\sum_{i-1}^{n} \hat{u}_i^2 \text{ , where } \hat{u} \text{ denotes the residuals } (y_i - \hat{y}); E(\hat{\sigma}^2) = \sigma^2.$$

Proof of the un-biasedness is left as an exercise (see p. 56 in the text).

Standard error of the OLS estimator

Substituting this into the variance of $\hat{\beta}_1$ and taking a square root gives the standard error of $\hat{\beta}_1$ as stated earlier.

$$se(\hat{\beta}_1) \equiv Var(\hat{\beta}_1)^{\frac{1}{2}} = \left[\frac{\sum_{i=1}^{n} \hat{u}_i{}^2}{(n-2)} * \left(\frac{1}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)\right]^{\frac{1}{2}} = \frac{\hat{\sigma}}{\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^{\frac{1}{2}}}$$

The standard error of an estimator with be eminently useful a bit later in the course when we discuss statistical inference, i.e., generating a confidence interval where the population parameter is expected to lie and testing whether its value is different from zero.

Conclusion

The topic of inference is postponed until chapter 4.

First we will generalize the simple linear regression model to include multiple explanatory variables.

y is uncorrelated with the residuals (optional)

$$2\sum_{i=1}^{n}\hat{u}_i(\hat{y}_i-\bar{y})=2\left[\sum_{i=1}^{n}\hat{u}_i\hat{y}_i-\bar{y}\sum_{i=1}^{n}\hat{u}_i\right]$$

Since the sum of residuals equals zero. Then we substitute conditional expectation of y.

$$2\left[\sum_{i=1}^{n}\hat{u}_i\hat{y}_i-\bar{y}\sum_{i=1}^{n}\hat{u}_i\right]=2\left[\sum_{i=1}^{n}\hat{u}_i(\bar{y}+\frac{s_{xy}}{s_x^2}(x_i-\bar{x}))-0\right]$$

$$2\left[\sum_{i=1}^{n}\hat{u}_i(\bar{y}+\frac{s_{xy}}{s_x^2}(x_i-\bar{x}))\right]=2\left[\bar{y}\sum_{i=1}^{n}\hat{u}_i+\frac{s_{xy}}{s_x^2}\sum_{i=1}^{n}\hat{u}_i(x_i-\bar{x})\right]$$

Y is uncorrelated with the residuals (concluded)

Again the sum of residuals is zero, so the first sum drops out.

$$2\left[0+\frac{s_{xy}}{s_x^2}\sum_{i=1}^{n}\hat{u}_i(x_i-\bar{x})\right]=2\left[\frac{s_{xy}}{s_x^2}\sum_{i=1}^{n}\hat{u}_i x_i-\bar{x}\frac{s_{xy}}{s_x^2}\sum_{i=1}^{n}\hat{u}_i\right]=2\left[\frac{s_{xy}}{s_x^2}\sum_{i=1}^{n}\hat{u}_i x_i\right]$$

Since the residuals from the estimated regression are independent, this sum is also zero.

Back.

Log-level models and percent change

$$\ln(y) = \beta_0 + \beta_1 x + u \rightarrow \frac{dln(y)}{dx} = \beta_1 = \frac{dln(y)}{dy}\frac{dy}{dx} = \frac{1}{y}\frac{dy}{dx}$$

$$\Leftrightarrow \beta_1 dx = \frac{dy}{y} = \frac{\%\Delta y}{100}, \text{and } dx = 1 \rightarrow 100 * \beta_1 = \%\Delta y.$$

Back.

Error term is mean zero

Say we had the following regression model with error term, $\tilde{u}$, that had a non-zero mean, $\bar{u}$, and that u_i was the de-meaned version of u tilde.

$$y_i = b_0 + \beta_1 x_i + \tilde{u}_i; u_i \equiv \tilde{u}_i - \bar{u} \Leftrightarrow \tilde{u}_i = u_i + \bar{u},$$

$$\text{so you could re} - \text{write}$$

$$y_i = b_0 + \beta_1 x_i + u_i + \bar{u},$$

$$\text{and combine } \bar{u} \text{ with } b_0, \text{calling the sum } \beta_0.$$

$$y_i = \beta_0 + \beta_1 x_i + u_i; \beta_0 \equiv b_0 + \bar{u}, \text{and } E(u_i) = 0.$$

Back.

x as a fixed regressor

See the “fixed in repeated samples” explanation on page 47.

It’s as if we took multiple samples of the same set of x observations, but each sample had different unobservable characteristics (u).

Practically speaking we can replace the denominator in the derivation (as they do in the text, p. 53) with a constant, as well as the “x” terms in the numerator.

Back.