Wooldridge, Introductory Econometrics, 5th ed.

Chapter 2: The simple regression model

Most of this course will be concerned with use of a regression model: a structure in which one or more explanatory variables are considered to generate an outcome variable, or dependent variable. We begin by considering the *simple regression* model, in which a single explanatory, or independent, variable is involved. We often speak of this as 'two-variable' regression, or 'Y on X regression'. Algebraically,

$$y_i = \beta_0 + \beta_1 x_i + u_i \tag{1}$$

is the relationship presumed to hold in the population for each observation i. The values of y are expected to lie on a straight line, depending on the corresponding values of x. Their values will differ from those predicted by that line by

the amount of the error term, or disturbance, u, which expresses the net effect of all factors other than x on the outcome y—that is, it reflects the assumption of *ceteris paribus*, or 'all else equal'.

We often speak of x as the 'regressor' in this relationship; less commonly we speak of y as the 'regressand.' The coefficients of the relationship, β_0 and β_1, are the regression parameters, to be estimated from a sample. They are presumed constant in the population, so that the effect of a one-unit change in x on y is assumed constant for all values of x.

As long as we include an intercept in the relationship, we can always assume that $E(u) = 0$, since a nonzero mean for u could be absorbed by the intercept term.

The crucial assumption in this regression model involves the relationship between x and u. We

consider x a random variable, as is u, and concern ourselves with the conditional distribution of u given x. If that distribution is equivalent to the unconditional distribution of u, then we can conclude that there is no relationship between x and u—which, as we will see, makes the estimation problem much more straightforward. To state this formally, we assume that

$$E(u \mid x) = E(u) = 0 \tag{2}$$

or that the u process has a **zero conditional mean**. This assumption states that the unobserved factors involved in the regression function are not related in any systematic manner to the observed factors.

For instance, consider a regression of individuals' hourly wage on the number of years of education they have completed. There are, of course, many factors influencing the hourly wage earned beyond the number of years of

formal schooling. In working with this regression function, we are assuming that the unobserved factors–excluded from the regression we estimate, and thus relegated to the u term–are not systematically related to years of formal schooling. This may not be a tenable assumption; we might consider "innate ability" as such a factor, and it is probably related to success in both the educational process and the workplace. Thus, innate ability–which we cannot measure without some proxies–may be positively correlated to the education variable, which would invalidate assumption (2).

The **population regression function**, given the zero conditional mean assumption, is

$$E(y \mid x) = \beta_0 + \beta_1 x_i \tag{3}$$

This allows us to separate y into two parts: the systematic part, related to x, and the unsystematic part, which is related to u. As long

as assumption (2) holds, those two components are independent in the statistical sense. Let us now derive the least squares estimates of the regression parameters.

Let $[(x_i, y_i) : i = 1, ..., n]$ denote a random sample of size n from the population, where y_i and x_i are presumed to obey the relation (1). The assumption (2) allows us to state that $E(u) = 0$, and given that assumption, that $Cov(x, u) = E(xu) = 0$, where $Cov(\cdot)$ denotes the covariance between the random variables. These assumptions can be written in terms of the regression error:

$$\begin{aligned} E\left(y_i - \beta_0 - \beta_1 x_i\right) &= 0 \\ E\left[x_i\left(y_i - \beta_0 - \beta_1 x_i\right)\right] &= 0 \end{aligned} \tag{4}$$

These two equations place two restrictions on the joint probability distribution of x and u. Since there are two unknown parameters to be

estimated, we might look upon these equations to provide solutions for those two parameters. We choose estimators b_0 and b_1 to solve the sample counterparts of these equations, making use of the principle of the method of moments:

$$\begin{aligned} n^{-1} \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i) &= 0 \qquad (5) \\ n^{-1} \sum_{i=1}^{n} x_i (y_i - b_0 - b_1 x_i) &= 0 \end{aligned}$$

the so-called normal equations of least squares. Why is this method said to be "least squares"? Because as we shall see, it is equivalent to minimizing the sum of squares of the regression residuals. How do we arrive at the solution? The first "normal equation" can be seen to be

$$b_0 = \bar{y} - b_1 \bar{x} \qquad (6)$$

where $\bar{y}$ and $\bar{x}$ are the sample averages of those variables. This implies that the regression line

passes through the point of means of the sample data. Substituting this solution into the second normal equation, we now have one equation in one unknown, b_1 :

$$
\begin{aligned}
0 &= \sum_{i=1}^{n} x_i \left(y_i - (\bar{y} - b_1 \bar{x}) - b_1 x_i\right) \\
\sum_{i=1}^{n} x_i \left(y_i - \bar{y}\right) &= b_1 \sum_{i=1}^{n} x_i \left(x_i - \bar{x}\right) \\
b_1 &= \frac{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2} \\
b_1 &= \frac{Cov(x, y)}{Var(x)} \qquad (7)
\end{aligned}
$$

where the slope estimate is merely the ratio of the sample covariance of the two variables to the variance of x, which, must be nonzero for the estimates to be computed. This merely implies that not all of the sample values of x

can take on the same value. There must be diversity in the observed values of x. These estimates–b_0 and b_1–are said to be the **ordinary least squares (OLS)** estimates of the regression parameters, since they can be derived by solving the least squares problem:

$$\min_{b_0,b_1} S = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2 \qquad (8)$$

Here we minimize the sum of squared residuals, or differences between the regression line and the values of y, by choosing b_0 and b_1. If we take the derivatives $\partial S/\partial b_0$ and $\partial S/\partial b_1$ and set the resulting first order conditions to zero, the two equations that result are exactly the OLS solutions for the estimated parameters shown above. The "least squares" estimates minimize the sum of squared residuals, in the sense that any other line drawn through the scatter of (x, y) points would yield a larger

sum of squared residuals. The OLS estimates provide the unique solution to this problem, and can always be computed if (i) $Var(x) > 0$ and (ii) $n \geq 2$. The estimated OLS regression line is then

$$\widehat{y}_i = b_0 + b_1 x_i \tag{9}$$

where the "hat" denotes the predicted value of y corresponding to that value of x. This is the **sample regression function (SRF)**, corresponding to the population regression function, or PRF (3). The population regression function is fixed, but unknown, in the population; the SRF is a function of the particular sample that we have used to derive it, and a different SRF will be forthcoming from a different sample. The primary interest in these estimates usually involves $b_1 = \partial y / \partial x = \Delta y / \Delta x$, the amount by which y is predicted to change from a unit change in the level of x. This slope

is often of economic interest, whereas the constant term in many regressions is devoid of economic meaning. In some cases, though, it has an economic interpretation.

For instance, a regression of major companies' CEO salaries on the firms' return on equity—a measure of economic performance—yields the regression estimates

$$\widehat{S} = 963.191 + 18.501r \tag{10}$$

where S is the CEO's annual salary, in thousands of 1990 dollars, and r is average return on equity over the prior three years, in per cent. This implies that a one percent increase in ROE over the past three years is worth \$18,501 to a CEO, on average. The average annual salary for the 209 CEOs in the sample is \$1.28 million, so the increment is about 1.4% of that average salary. The SRF can also be used to predict what a CEO will

earn for any level of ROE; points on the estimated regression function are such predictions.

Mechanics of OLS

Some algebraic properties of the OLS regression line:

(1) The sum (and average) of the OLS residuals is zero:

$$\sum_{i=1}^{n} e_i = 0 \tag{11}$$

which follows from the first normal equation, which specifies that the estimated regression line goes through the point of means $(\bar{x}, \bar{y})$, so that the mean residual must be zero.

(2) By construction, the sample covariance between the OLS residuals and the regressor is zero:

$$Cov(e, x) = \sum_{i=1}^{n} x_i e_i = 0 \tag{12}$$

This is not an assumption, but follows directly from the second normal equation. The estimated coefficients, which give rise to the residuals, are chosen to make it so.

(3) Each value of the dependent variable may be written in terms of its prediction and its error, or regression residual:

$$y_i = \widehat{y}_i + e_i$$

so that OLS decomposes each y_i into two parts: a fitted value, and a residual. Property (3) also implies that $Cov(e, \widehat{y}) = 0$, since $\widehat{y}$ is a linear transformation of x, and linear transformations have linear effects on covariances. Thus, the fitted values and residuals are uncorrelated in the sample. Taking this property and applying it to the entire sample, we define

$$
\begin{aligned}
SST &= \sum_{i=1}^{n} (y_i - \bar{y})^2 \\
SSE &= \sum_{i=1}^{n} (\widehat{y}_i - \bar{y})^2 \\
SSR &= \sum_{i=1}^{n} e_i^2
\end{aligned}
$$

as the Total sum of squares, Explained sum of squares, and Residual sum of squares, respectively. Note that SST expresses the total variation in y around its mean (and we do not strive to "explain" its mean; only how it varies about its mean).

The second quantity, SSE, expresses the variation of the predicted values of y around the mean value of y (and it is trivial to show that $\widehat{y}$ has the same mean as y).

The third quantity, SSR, is the same as the least squares criterion S from (8). (Note that

some textbooks interchange the definitions of SSE and SSR, since both "explained" and "error" start with E, and "regression" and "residual" start with R). Given these sums of squares, we can generalize the decomposition mentioned above into

$$SST = SSE + SSR \tag{13}$$

or, the total variation in y may be divided into that **explained** and that **unexplained**, i.e. left in the residual category. To prove the validity of (13), note that

$$
\begin{aligned}
\sum_{i=1}^{n} (y_i - \bar{y})^2 &= \sum_{i=1}^{n} ((y_i - \widehat{y}_i) + (\widehat{y}_i - \bar{y}))^2 \\
&= \sum_{i=1}^{n} [e_i + (\widehat{y}_i - \bar{y})]^2 \\
&= \sum_{i=1}^{n} e_i^2 + 2 \sum_{i=1}^{n} e_i (\widehat{y}_i - \bar{y}) + \\
&\quad \sum_{i=1}^{n} (\widehat{y}_i - \bar{y})^2 \\
SST &= SSR + SSE
\end{aligned}
$$

given that the middle term in this expression is equal to zero. But this term is the sample covariance of e and y, given a zero mean for e, and by (12) we have established that this is zero.

How good a job does this SRF do? Does the regression function explain a great deal of the variation of y, or not very much? That can

now be answered by making use of these sums of squares:

$$R^2 = [r_{xy}]^2 = \frac{SSE}{SST} = 1 - \frac{SSR}{SST}$$

The R^2 measure (sometimes termed the coefficient of determination) expresses the percent of variation in y around its mean "explained" by the regression function. It is an r, or simple correlation coefficient, squared, in this case of simple regression on a single x variable. Since the correlation between two variables ranges between -1 and +1, the squared correlation ranges between 0 and 1. In that sense, R^2 is like a batting average.

In the case where $R^2 = 0$, the model we have built fails to explain any of the variation in the y values around their mean–unlikely, but it is certainly possible to have a very low value of R^2. In the case where $R^2 = 1$, all of the points lie on the SRF. That is unlikely when $n > 2$, but

it may be the case that all points lie close to the line, in which case R^2 will approach 1. We cannot make any statistical judgment based directly on R^2, or even say that a model with a higher R^2 and the same dependent variable is necessarily a better model; but other things equal, a higher R^2 will be forthcoming from a model that captures more of $y's$ behavior.

In cross-sectional analyses, where we are trying to understand the idiosyncracies of individual behavior, very low R^2 values are common, and do not necessarily denote a failure to build a useful model. In time-series analyses, much higher R^2 values are often commonly observed, as many economic and financial variables move together over time.

Important issues in evaluating applied work: how do the quantities we have estimated change when the units of measurement are changed?

In the estimated model of CEO salaries, since the y variable was measured in thousands of dollars, the intercept and slope coefficient refer to those units as well. If we measured salaries in dollars, the intercept and slope would be multiplied by 1000, but nothing would change. The correlation between y and x is not affected by linear transformations, so we would not alter the R^2 of this equation by changing its units of measurement. Likewise, if ROE was measured in decimals rather than per cent, it would merely change the units of measurement of the slope coefficient. Dividing r by 100 would cause the slope to be multiplied by 100. In the original (10), with r in percent, the slope is 18.501 (thousands of dollars per one unit change in r). If we expressed r in decimal form, the slope would be 1850.1. A change in r from 0.10 to 0.11 – a one per cent increase in ROE–would be associated with a change in salary of (0.01)(1850.1)=18.501 thousand

dollars. Again, the correlation between salary and ROE would not be altered. This also applies for a transformation such as $F = 32 + \frac{9}{5}C$; it would not matter whether we viewed temperature in degrees F or degrees C as a causal factor in estimating the demand for heating oil, since the correlation between the dependent variable and temperature would be unchanged by switching from Fahrenheit to Celsius degrees.

Functional form

Simple linear regression would seem to be a workable tool if we have a presumed linear relationship between y and x, but what if theory suggests that the relation should be nonlinear? It turns out that the "linearity" of regression refers to y being expressed as a linear function of x—but neither y nor x need be the "raw data" of our analysis. For instance, regressing

y on t (a time trend) would allow us to analyse a linear trend, or constant growth, in the data. What if we expect the data to exhibit exponential growth, as would population, or sums earning compound interest? If the underlying model is

$$\begin{aligned} y &= A \exp(rt) \\ \log y &= \log A + rt \\ y^* &= A^* + rt \end{aligned} \tag{14}$$

so that the "single-log" transformation may be used to express a constant-growth relationship, in which r is the regression slope coefficient that directly estimates $\partial \log y / \partial t$. Likewise, the "double-log" transformation can be used to express a constant-elasticity relationship, such as that of a Cobb-Douglas function:

$$\begin{aligned} y &= Ax^{\alpha} \\ \log y &= \log A + \alpha \log x \\ y^* &= A^* + \alpha x^* \end{aligned} \tag{15}$$

In this context, the slope coefficient α is an estimate of the elasticity of y with respect to x, given that $\eta_{y,x} = \partial \log y / \partial \log x$ by the definition of elasticity. The original equation is nonlinear, but the transformed equation is a linear function which may be estimated by OLS regression.

Likewise, a model in which y is thought to depend on $1/x$ (the reciprocal model) may be estimated by linear regression by just defining a new variable, z, equal to $1/x$ (presuming $x > 0$). That model has an interesting interpretation if you work out its algebra.

We often use a polynomial form to allow for nonlinearities in a regression relationship. For instance, rather than including only x as a regressor, we may include x and x^2. Although this relationship is linear in the parameters, it implies that $\frac{\partial Y}{\partial x} = \beta + 2\gamma x$, so that the effect

of x on Y now depends on the level of x for that observation, rather than being a constant factor.

Properties of OLS estimators

Now let us consider the properties of the regression estimators we have derived, considering b_0 and b_1 as estimators of their respective population quantities. To establish the unbiasedness of these estimators, we must make several assumptions:

Proposition 1 *SLR1: in the population, the dependent variable y is related to the independent variable x and the error u as*

$$y = \beta_0 + \beta_1 x + u \tag{16}$$

Proposition 2 *SLR2: we can estimate the population parameters from a sample of size n, $\{(x_i, y_i), i = 1, ..., n\}$.*

Proposition 3 *SLR3: the error process has a zero conditional mean:*

$$E\left(u \mid x\right) = 0. \tag{17}$$

Proposition 4 *SLR4: the independent variable x has a positive variance:*

$$\left(n-1\right)^{-1} \sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2 > 0. \tag{18}$$

Given these four assumptions, we may proceed, considering the intercept and slope estimators as random variables. For the slope estimator; we may express the estimator in terms of population coefficients and errors:

$$b_1 = \frac{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2} = \frac{\sum_{i=1}^{n} \left(x_i - \bar{x}\right) y_i}{s_x^2} \tag{19}$$

where we have defined s_x^2 as the total variation in x (not the variance of x). Substituting, we can write the slope estimator as:

$$b_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})\, y_i}{s_x^2}$$

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

$$= \frac{\beta_0 \sum_{i=1}^{n} (x_i - \bar{x}) + \beta_1 \sum_{i=1}^{n} (x_i - \bar{x})\, x_i + \sum_{i=1}^{n} (x_i - \bar{x})\, u_i}{s_x^2} \tag{20}$$

We can show that the first term in the numerator is algebraically zero, given that the deviations around the mean sum to zero. The second term can be written as $\sum_{i=1}^{n} (x_i - \bar{x})^2 = s_x^2$, so that the second term is merely β_1 when divided by s_x^2. Thus this expression can be rewritten as:

$$b_1 = \beta_1 + \frac{1}{s_x^2} \sum_{i=1}^{n} (x_i - \bar{x})\, u_i$$

showing that any randomness in the estimates of b_1 is derived from the errors in the sample, weighted by the deviations of their respective

x values. Given the assumed independence of the distributions of x and u implied by (17), this expression implies that:

$$E\left(b_1\right) = \beta_1,$$

or that b_1 is an unbiased estimate of β_1, given the propositions above. The four propositions listed above are all crucial for this result, but the key assumption is the independence of x and u.

We are also concerned about the precision of the OLS estimators. To derive an estimator of the precision, we must add an assumption on the distribution of the error u :

Proposition 5 *SLR5: (***homoskedasticity***):*

$$Var\left(u \mid x\right) = Var(u) = \sigma^2.$$

This assumption states that the variance of the error term is constant over the population, and

thus within the sample. Given (17), the conditional variance is also the unconditional variance. The errors are considered drawn from a fixed distribution, with a mean of zero and a constant variance of σ^2. If this assumption is violated, we have the condition of heteroskedasticity, which will often involve the magnitude of the error variance relating to the magnitude of x, or to some other measurable factor.

Given this additional assumption, but no further assumptions on the nature of the distribution of u, we may demonstrate that:

$$Var\left(b_1\right) = \frac{\sigma^2}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2} = \frac{\sigma^2}{s_x^2} \tag{21}$$

so that the precision of our estimate of the slope is dependent upon the overall error variance, and is inversely related to the variation in the x variable. The magnitude of x does not matter, but its variability in the sample does

matter. If we are conducting a controlled experiment (quite unlikely in economic analysis) we would want to choose widely spread values of x to generate the most precise estimate of $\partial y/\partial x$.

We can likewise prove that b_0 is an unbiased estimator of the population intercept, with sampling variance:

$$Var\left(b_0\right) = n^{-1}\frac{\sigma^2 \sum_{i=1}^{n} x_i^2}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2} = \frac{\sigma^2 \sum_{i=1}^{n} x_i^2}{n s_x^2} \tag{22}$$

so that the precision of the intercept depends, as well, upon the sample size, and the magnitude of the x values. These formulas for the sampling variances will be invalid in the presence of heteroskedasticity–that is, when proposition SLR5 is violated.

These formulas are not operational, since they include the unknown parameter σ^2. To calculate estimates of the variances, we must first

replace σ^2 with a consistent estimate, s^2, derives from the least squares residuals:

$$e_i = y_i - b_0 - b_1 x_i, \;\; i = 1, ..., n \qquad (23)$$

We cannot observe the error u_i for a given observation, but we can generate a consistent estimate of the ith observation's error with the ith observation's least squares **residual**, $\widehat{u}_i$. Likewise, a sample quantity corresponding to the population variance σ^2 can be derived from the residuals:

$$s^2 = \frac{1}{(n-2)} \sum_{i=1}^{n} e_i^2 = \frac{SSR}{(n-2)} \qquad (24)$$

where the numerator is just the least squares criterion, SSR, divided by the appropriate degrees of freedom. Here, two degrees of freedom are lost, since each residual is calculated by replacing two population coefficients with their sample counterparts. This now makes it possible to generate the estimated variances

and, more usefully, the **estimated standard error** of the regression slope:

$$s_{b_1} = \frac{s}{s_x}$$

where s is the standard deviation, or standard error, of the disturbance process (that is, $\sqrt{s^2}$), and s_x is $\sqrt{s_x^2}$. It is this estimated standard error that will be displayed on the computer printout when you run a regression, and used to construct confidence intervals and hypothesis tests about the slope coefficient. We can calculate the estimated standard error of the intercept term by the same means.

Regression through the origin

We could also consider a special case of the model above where we impose a constraint that $\beta_0 = 0$, so that y is taken to be proportional to x. This will often be inappropriate; it is generally more sensible to let the data calculate the appropriate intercept term, and reestimate the model subject to that constraint only if that is a reasonable course of action. Otherwise, the resulting estimate of the slope coefficient will be biased. Unless theory suggests that a strictly proportional relationship is appropriate, the intercept should be included in the model.