A Brief Guide to Linear Regression

A common statistical tool used to uncover empirical relationships between variables in economics is linear regression. A linear regression expresses a dependent variable as a linear combination of explanatory variables and an error term. The error term represents the contribution to the dependent variable of any omitted or unobserved variables. In section I, we will review some fundamental concepts which will facilitate our discussion of the linear regression model. Then, in section II, we will consider how the marginal effects (coefficients) of an explanatory variable upon the dependent variable may be estimated in a linear regression model. Finally, in section III, we will outline a common method for conducting inference within a linear regression model. We will be able to quantify the degree of confidence we place in our coefficient estimates.

I Preliminaries

Before beginning any statistical investigation, we formulate the question we would like to answer. For example:

1. does greater trade openness lead to greater economic growth for a country?
2. does higher educational attainment lead to higher real incomes?
3. does happiness increase with real income?

The question we ask necessarily restricts the kind of objects under consideration. In the case of the first example question, we are considering countries. In the cases of the second and third example questions, we need to narrow further. They could be questions about individuals or countries (if they implicitly refer to the behavior of average quantities). The complete universe of objects under consideration is known as the **population**.

Many populations may be described by their moments – the average values of population characteristics or **expectations**. The most important moments for our purposes are: the **mean**, the **variance**, and the **covariance**:

- A population mean is merely the average value of a raw population characteristic (*e.g.*, the average level of education across individuals in a nation). Denoting the characteristic by the random variable x, the population mean is written as $E(x)$ and is read as 'the expected value of x.' Note that although x is a random variable (*viz.*, its exact value for an object is not specified *ex ante*), its expectations are not. For example, the population mean is a **parameter** – a fixed value. Parameters may be known or unknown.
- The variance is the average squared distance of a variable from its mean. It is a measure of the spread of a characteristic quantified by the variable. Again, denoting the characteristic by the random variable x, its variance is written as $Var(x) = E\left\{[x - E(x)]^2\right\}$.

- As its name implies, the covariance quantifies how two characteristics/variables are linearly related (*viz.*, how they move together or covary). It is defined as the expected value of the product of two random variables minus the product of their respective means. Denoting the two characteristics by the random variables x and y, it is written as $Cov(x, y) = E\{[x - E(x)][y - E(y)]\} = E(xy) - E(x)E(y)$.

Due to cost and measurement problems, it is oftentimes infeasible to gather all of the information necessary to answer a question definitively about a population. However, we can still make some progress by collecting data on a subset of the population that is carefully chosen – a **sample**. The sample consists of **observations**. An observation is a data record on one unit or object under consideration. Ideally, we choose the sample so that the objects within it generally occur in the same proportion as they do in the population – the sample is representative.[1] The sample represents the population in miniature. Alternatively, we might think of the population as the sample we would obtain with an unlimited number of observations.

With a representative sample, the **analogy principle** states that we may learn something about population moments by forming their sample analogs. For example, an estimate of the population mean for a characteristic may be derived by calculating the sample mean for the same characteristic. For a characteristic quantified by x, we write:

$$\begin{array}{c} \text{Sample} \rightarrow \text{Population} \\ \widehat{E}(x) = \frac{1}{N}\sum_{i=1}^{N} x_i \rightarrow E(x) \end{array}$$

where the hat denotes the sample analog, N denotes the sample size (number of observations), and i indexes observations.
Nota Bene: It is important to distinguish between an **estimator** and an **estimate**. An estimator is a known function of random variables, and is thus itself a random variable. As a random variable, an estimator can be characterized by its moments, similar to other random variables. An estimate is merely a particular value taken by an estimator; it is determined by the specific realizations of the random variables which constitute the estimator.

II Estimation

Suppose that we have a sample of data on a single explanatory variable (x) and a single dependent variable (y) – (x_i, y_i) for $i = 1, N$. Here, i indexes the observation in the sample, which contains N data points. The sample is representative of some underlying population of interest. In the case of a single explanatory variable, the linear regression equation for an observation may be written as:

$$y_i = x_i\beta + \varepsilon_i,$$

where ε is the error term (the contribution of omitted or unobserved variables to y). This is the simplest possible linear regression model. The coefficient β is the effect of x upon

[1] Random sampling is one means of guaranteeing that the sample is representative.

y, which we would like to know. It is an unknown population parameter. Notice that β is assumed to be constant across observations.

Since we do not observe ε, we cannot solve for β without some additional assumptions. Specifically, we need assumptions which will allow us to estimate β solely as a function of observables (x and y). Some straightforward algebra shows us a way forward. First, we multiply the linear regression equation for each observation by x_i:

$$x_i y_i = x_i^2 \beta + x_i \varepsilon_i.$$

Second, the constancy of β means that we can sum across all observations in the sample and factor out β:

$$\begin{aligned}
\sum_{i=1}^{N} x_i y_i &= \sum_{i=1}^{N} x_i^2 \beta + \sum_{i=1}^{N} x_i \varepsilon_i \Rightarrow \\
\beta &= \frac{\sum_{i=1}^{N} x_i y_i}{\sum_{i=1}^{N} x_i^2} - \frac{\sum_{i=1}^{N} x_i \varepsilon_i}{\sum_{i=1}^{N} x_i^2} \\
&= \frac{\frac{1}{N}\left(\sum_{i=1}^{N} x_i y_i\right)}{\frac{1}{N}\left(\sum_{i=1}^{N} x_i^2\right)} - \frac{\frac{1}{N}\left(\sum_{i=1}^{N} x_i \varepsilon_i\right)}{\frac{1}{N}\left(\sum_{i=1}^{N} x_i^2\right)}.
\end{aligned}$$

We can now see that β is equal to the sum of two terms. The first term is the average product of x and y divided by the average square of x; it is only a function of observables. The second term is the average product of x and ε divided by the average square of x; it is a function of both observables and unobservables.

If the sample average product of x and ε is usually zero, then we can ignore the second term and generally be accurate in our estimate of β. Thus, under the assumption that x and ε are *unrelated on average* in this sense, we can estimate β by:

$$\hat{\beta} = \frac{\frac{1}{N}\left(\sum_{i=1}^{N} x_i y_i\right)}{\frac{1}{N}\left(\sum_{i=1}^{N} x_i^2\right)} = \frac{\widehat{E}(xy)}{\widehat{E}(x^2)}.$$

This estimator of β (denoted by β with a hat over it) is known as the **ordinary least squares (OLS)** estimator.[2]

We can readily generalize from a single explanatory variable to multiple explanatory variables, where the linear regression equation takes the form:

$$y_i = x_{1,i}\beta_1 + x_{2,i}\beta_2 + x_{3,i}\beta_3 + \cdots + x_{K,i}\beta_K + \varepsilon_i.$$

[2]It is a least squares estimator because it may also be derived through an explicit optimization problem – choose the value of β which minimizes the sum of squared residuals (another name for the errors), $\sum_{i=1}^{N}(y_i - x_i\beta)^2$. According to this objective, β is chosen to explain as much variation of y as possible given the variation in x. In this manner, the resulting sum of squared residuals is the least possible. The adjective ordinary refers to how standard errors are constructed for this estimator; we will consider this later.

Here, there are K explanatory variables, with the variables and their coefficients suitably subscripted. Notice how this allows for an intercept term to be easily introduced – set $x_{1,i} = 1$ for all i. We will not derive explicit formulae for the multiple linear regression coefficients, other than for the special case of an equation with an intercept and slope.

Consider the case with two explanatory variables, where $x_{1,i} = 1$ for all i and $x_{2,i}$ is denoted by x_i. The linear regression equation may then be written as:

$$y_i = \alpha + x_i\beta + \varepsilon_i,$$

where $\alpha = \beta_1$ and $\beta = \beta_2$, from the earlier notation. This regression model will be our ***baseline model***, from which we will build the rest of our econometric intuition. As before, the coefficients (α and β) and the error term are unobserved. The intercept α represents the average value of y after controlling for x. The slope β represents the effect of x upon y, controlling for y's average value. As before, we would like to construct estimators of α and β which are only functions of the observables (x, y). Again, we will require some additional assumptions to eliminate ε from our calculations. Summing across all observations in the sample and factoring out α and β, we have that:

$$\begin{aligned}
\sum_{i=1}^{N} y_i &= \sum_{i=1}^{N} \alpha + \sum_{i=1}^{N} x_i\beta + \sum_{i=1}^{N} \varepsilon_i \Rightarrow \\
\sum_{i=1}^{N} y_i &= \alpha N + \beta \sum_{i=1}^{N} x_i + \sum_{i=1}^{N} \varepsilon_i \Rightarrow \\
\left(\frac{1}{N}\sum_{i=1}^{N} y_i\right) &= \alpha + \beta\left(\frac{1}{N}\sum_{i=1}^{N} x_i\right) + \left(\frac{1}{N}\sum_{i=1}^{N} \varepsilon_i\right) \\
\bar{y} &= \alpha + \bar{x}\beta + \bar{\varepsilon},
\end{aligned}$$

where overbars denote sample means. Then, we know that:

$$\alpha = \bar{y} - \bar{x}\beta - \bar{\varepsilon}.$$

Substituting this expression for α into our original linear regression equation, we have that:

$$\begin{aligned}
y_i &= \bar{y} - \bar{x}\beta - \bar{\varepsilon} + x_i\beta + \varepsilon_i \Rightarrow \\
(y_i - \bar{y}) &= (x_i - \bar{x})\,\beta + (\varepsilon_i - \bar{\varepsilon})\,.
\end{aligned}$$

The regression equation now resembles the linear regression equation with only a slope. By de-meaning the variables (subtracting their sample means from each observation), we have transformed the regression into a familiar form. Then, we know that:

$$\begin{aligned}
\beta &= \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N}(x_i - \bar{x})^2} - \frac{\sum_{i=1}^{N}(x_i - \bar{x})(\varepsilon_i - \bar{\varepsilon})}{\sum_{i=1}^{N}(x_i - \bar{x})^2} \Rightarrow \\
\hat{\beta} &= \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N}(x_i - \bar{x})^2} = \frac{\widehat{Cov}(x, y)}{\widehat{Var}(x)},
\end{aligned}$$

where we have invoked the OLS assumption that x and ε are *unrelated on average*.[3] Here, this means that they they have zero covariance or are **uncorrelated**. Using $\hat{\beta}$, an OLS estimator of α may be calculated as:

$$\hat{\alpha} = \bar{y} - \bar{x}\hat{\beta},$$

if the sample average of ε is close to $E(\varepsilon) = 0$, as we assumed. Here, the assumption that $\bar{\varepsilon} = 0$, required to estimate the intercept, is analogous to the assumption that x and ε are unrelated on average which is required to estimate the slope.

In summary, accurately estimating a linear regression for a dependent variable requires that the following assumptions hold:

1. The coefficients (*viz.*, the parameters) are constant across observations in the sample.

2. The explanatory variables and the error term are *unrelated on average*. In a regression model with an intercept (our baseline model), this is equivalent to the assumption that they have zero covariance $\rightarrow Cov(x, \varepsilon) = E(x\varepsilon) = 0$. *Viz.*, they are **uncorrelated**.

   (a) Since the error term contains any omitted or unobserved explanatory variables, we are requiring that the included explanatory variables be uncorrelated with any omitted variables that are relevant.

III Inference

When interpreting the evidence regarding the relationship between y and x from the linear regression, we need to take account of the inherent uncertainty of our coefficient estimator; $\hat{\beta}$ is not β, since it is derived from a sample. Even if the sample *is* the population of interest, measurement error in the dependent variable y may lead to some deviation between $\hat{\beta}$ and β. Thus, we would like to get a quantitative sense of the degree of uncertainty we have regarding how well the coefficient estimate matches the true parameter.

If we think of x as being fixed or non-stochastic (non-random), the only source of randomness in a linear regression model is ε, the error term. Then, it is obvious that $\hat{\beta}$ is a random variable, since it is a function of y which is a function of ε. As a random variable, $\hat{\beta}$ may be characterized by its moments. Given the assumptions of the OLS estimator, it can be shown that:

$$E\left(\hat{\beta}\right) = \beta,$$

where we have used the property that the expectation of a constant is the constant. Thinking of x as non-stochastic, the only random variable inside the estimator $\hat{\beta}$ is ε, which has expectation zero under the OLS assumptions.

[3] The hat over the Cov and Var indicates that these are the sample analogs of the population moments.

III.1 Variance

A natural measure of the degree of uncertainty surrounding a random variable is its variance – the average squared distance of the variable from its expected value. In our baseline model (the regression model with an intercept), the variance of $\hat{\beta}$ is given by:

$$\begin{aligned}
Var\left(\hat{\beta}\right) &= E\left(\hat{\beta}-\beta\right)^2 \\
&= E\left(\hat{\beta}^2\right)-\beta^2 \\
&= \frac{E\left(\left[\sum_{i=1}^{N}\left(x_i-\bar{x}\right)\left(y_i-\bar{y}\right)\right]^2\right)}{\left[\sum_{i=1}^{N}\left(x_i-\bar{x}\right)^2\right]^2}-\beta^2 \\
&= \frac{E\left(\left[\sum_{i=1}^{N}\left(x_i-\bar{x}\right)^2\beta+\sum_{i=1}^{N}\left(x_i-\bar{x}\right)\left(\varepsilon_i-\bar{\varepsilon}\right)\right]^2\right)}{\left[\sum_{i=1}^{N}\left(x_i-\bar{x}\right)^2\right]^2}-\beta^2 \\
&= \frac{\beta^2\left[\sum_{i=1}^{N}\left(x_i-\bar{x}\right)^2\right]^2+2\beta\left[\sum_{i=1}^{N}\left(x_i-\bar{x}\right)^2\right]\left[\sum_{i=1}^{N}\left(x_i-\bar{x}\right)E\left(\varepsilon_i-\bar{\varepsilon}\right)\right]}{\left[\sum_{i=1}^{N}\left(x_i-\bar{x}\right)^2\right]^2} \\
&\quad +\frac{E\left(\left[\sum_{i=1}^{N}\left(x_i-\bar{x}\right)\left(\varepsilon_i-\bar{\varepsilon}\right)\right]^2\right)}{\left[\sum_{i=1}^{N}\left(x_i-\bar{x}\right)^2\right]^2}-\beta^2 \\
&= \frac{E\left(\left[\sum_{i=1}^{N}\left(x_i-\bar{x}\right)\left(\varepsilon_i-\bar{\varepsilon}\right)\right]^2\right)}{\left[\sum_{i=1}^{N}\left(x_i-\bar{x}\right)^2\right]^2} \\
&= \frac{\left(\sum_{i=1}^{N}\sum_{j=1}^{N}\left(x_i-\bar{x}\right)\left(x_j-\bar{x}\right)E\left[\left(\varepsilon_i-\bar{\varepsilon}\right)\left(\varepsilon_j-\bar{\varepsilon}\right)\right]\right)}{\left[\sum_{i=1}^{N}\left(x_i-\bar{x}\right)^2\right]^2},
\end{aligned}$$

where we have used: the definition of y as the linear combination of x and ε, the assumption that ε is the only source of randomness, and the assumption that $E(\varepsilon)=0$. However, to get any farther, we need to make some additional assumptions regarding the behavior of the unobservable error term, ε. Specifically, we require that:

1. ε_i is uncorrelated with ε_{-i}, where the subscript $-i$ denotes any observation other than $i \Rightarrow Cov(\varepsilon_i,\varepsilon_{-i})=0$ for all i. In other words, the errors are not correlated across observations. Since $E(\varepsilon)=0$ by assumption, this also implies that $E(\varepsilon_i,\varepsilon_{-i})=0$ for all i.

2. ε has the same variance across observations $\Rightarrow Var(\varepsilon_i)=\sigma^2$ for all i.

With the uncorrelatedness and homoskedasticity (identical variances) of ε, the formula for the variance of $\hat{\beta}$ simplifies further:

$$\begin{aligned}
Var\left(\hat{\beta}\right) &= \frac{\left[\sum_{i=1}^{N}\sum_{j=1}^{N}\left(x_i-\bar{x}\right)\left(x_j-\bar{x}\right)E\left(\varepsilon_i\varepsilon_j\right)\right]}{\left[\sum_{i=1}^{N}\left(x_i-\bar{x}\right)^2\right]^2} \\
&= \frac{\left[\sum_{i=1}^{N}\left(x_i-\bar{x}\right)^2 E\left(\varepsilon_i^2\right)\right]}{\left[\sum_{i=1}^{N}\left(x_i-\bar{x}\right)^2\right]^2} = \frac{\left[\sum_{i=1}^{N}\left(x_i-\bar{x}\right)^2 Var\left(\varepsilon_i\right)\right]}{\left[\sum_{i=1}^{N}\left(x_i-\bar{x}\right)^2\right]^2} \\
&= \frac{\left[\sum_{i=1}^{N}\left(x_i-\bar{x}\right)^2 \sigma^2\right]}{\left[\sum_{i=1}^{N}\left(x_i-\bar{x}\right)^2\right]^2} = \frac{\sigma^2\left[\sum_{i=1}^{N}\left(x_i-\bar{x}\right)^2\right]}{\left[\sum_{i=1}^{N}\left(x_i-\bar{x}\right)^2\right]^2} \\
&= \frac{\sigma^2}{\left[\sum_{i=1}^{N}\left(x_i-\bar{x}\right)^2\right]} \\
&= \frac{\sigma^2}{\widehat{Var}\left(x_i\right)}.
\end{aligned}$$

By the analogy principle, we can estimate σ^2 with:

$$\begin{aligned}
\widehat{\sigma}^2 &= \widehat{Var}\left(\varepsilon\right) \\
&= \frac{1}{N}\sum_{i=1}^{N}\left(y_i - x_i\hat{\beta}\right)^2 \Rightarrow \\
\widehat{Var}\left(\hat{\beta}\right) &= \frac{\widehat{\sigma}^2}{\widehat{Var}\left(x_i\right)}.
\end{aligned}$$

If the assumptions we used to derive the formulae are correct, then our estimators will give us accurate estimates.[4] The **standard error** of $\hat{\beta}$ is defined to be the square root of the estimated variance:

$$\widehat{SE}\left(\hat{\beta}\right) = \sqrt{\widehat{Var}\left(\hat{\beta}\right)}.$$

It is easier to interpret than the variance, since it is the average distance of $\hat{\beta}$ from its mean instead of the average squared distance.

III.2 Confidence

We now have a measure of the spread associated with our estimator. If we take the ratio of our estimate of β to its standard error (as calculated above), then we have the

[4]The estimator for σ^2 here is not the standard OLS estimator of σ^2, but rather the maximum likelihood estimator of σ^2. They are typically extremely close. This is a technical detail which you need not worry about for our course.

t-statistic associated with $\hat{\beta}$:[5]

$$t_{\hat{\beta}} = \frac{\hat{\beta}}{\widehat{SE}\left(\hat{\beta}\right)}.$$

Under some additional assumptions about ε, the t-statistic tells us whether or not it is reasonable to argue that the population β is actually zero, given the information we have from the sample in $\hat{\beta}$. The **rule-of-thumb** is that:

$$t_{\hat{\beta}} > 2 \Rightarrow \text{With 95\% confidence, the true } \beta \neq 0.$$

For the rule-of-thumb to be exact, we require that ε be normally distributed (follows a bell-curve). Even without a normally distributed error term, the rule-of-thumb is approximately correct.[6] If a t-statistic fulfills this criterion, the corresponding coefficient estimate is said to be significant at the 5% level.

In words, 95% of the time (with various samples), the estimator $\hat{\beta}$ will have a t-statistic that is less than 2 if the true $\beta = 0$. Thus, if the t-statistic is greater than 2, then we will mistakenly infer that $\beta \neq 0$ when in fact $\beta = 0$ only 5% of the time. The 5% chance of an error is typically deemed acceptable for most applications, leading to our rule-of-thumb for inference.

III.3 Predictability and Fit

A common measure of the explanatory power of a linear regression is **R-squared**, denoted R^2.[7] It is the proportion of the total variation in the dependent variable for the sample which is predicted by an associated linear regression. Specifically, it is:

$$R^2 = \frac{\sum_{i=1}^{N}\left(\widehat{y}_i - \bar{y}\right)^2}{\sum_{i=1}^{N}\left(y_i - \bar{y}\right)^2} = 1 - \frac{\sum_{i=1}^{N}\left(y_i - \widehat{y}_i\right)^2}{\sum_{i=1}^{N}\left(y_i - \bar{y}\right)^2} \in [0,1],$$

where $\widehat{y}_i$ is the predicted or fitted value from the linear regression for observation i. Notice that the R-squared is always bounded between zero and one – zero indicates that the linear regression can predict *none* of the variation in y, while one indicates that the linear regression can *perfectly* predict y.[8]

Because it is calculated from the sample, a regression's R-squared is only a measure of the in-sample fit. A high R-squared does not imply that the regression will predict well for a new sample or additional observations (*viz.*, the out-of-sample fit may not be high).

---

[5] It is called a t-statistic because its exact statistical interpretation is related to Student's t distribution.

[6] The proof relies upon central limit theorems, which are beyond the scope of our course.

[7] It is also sometimes known as the multiple coefficient of determination or multiple correlation coefficient (squared).

[8] The implicit benchmark against which the linear regression is judged, is the sample mean of y. If the linear regression does a better job of prediction than the sample mean, then the R-squared will be strictly positive. For linear regression models which include an intercept term, this will always be the case.

III.4 Threats to Valid Inference

Significant t-statistics and a high R-squared suggest that the related x variables are important determinants of the y variable. However, since linear regression essentially relies upon correlation to make inference, there are several *caveats*:

1. If the assumptions underlying the linear regression are false, then the resulting coefficient estimates are not accurate.

   (a) **Constancy of β** – If the population parameter is not constant across observations, then $\hat{\beta}$ may only be interpreted as the average effect of x upon y in the sample. A richer statistical model will be required to infer anything robust about the population.

   (b) **Uncorrelatedness of x and ε** – There are three main ways in which this may fail in a linear regression:

      i. *Reverse causation* – If y causes x, then the coefficient upon x may appear to indicate that x is a significant determinant of y in the linear regression, even though x is not causal for y.

      ii. *Simultaneity* – If two variables are jointly determined, then the simple interpretation that x is a significant determinant of y will not be appropriate. With simultaneity, the dependent and explanatory variables are part of a feedback process. The classical example in economics is the price and quantity transacted of a good in a decentralized market.

      iii. *Omitted or unobserved variables* – If x and y are both caused by some third variable which is not included in the linear regression, then it may appear that x is a significant determinant of y, even though it is not. The included explanatory variable attempts to account for the variation in y through its relationship with the important, omitted third variable.

2. Even if the assumptions underlying the linear regression are true, the t-statistics will not be accurate if the assumptions underlying the variance derivation are false.

   (a) **Uncorrelatedness of ε_i with ε_{-i}** – If the errors are correlated across observation in the sample, then the regression exhibits *autocorrelation* or *serial correlation*. The calculated standard errors will not be accurate.

   (b) **Constant variance of ε across observation (homoskedasticity)** – If the variance of the error term is different across observations, then the regression exhibits *heteroskedasticity*. Again, the calculated standard errors will not be accurate.

When interpreting the results of a linear regression, we must be ever-vigilant to consider these possible threats. By far, the most serious are the threats of the first type, as they mean that the linear regression is incorrect (*i.e.*, garbage in, garbage out). The

threats of the second type are less worrying, as there are statistical methods which may ameliorate them. However, the exact methods lie beyond the scope of this course. In general, we will only concern ourselves with thinking about threats of the first type when interpreting regression results.