Michele Pellizzari

(University of Geneva)

Univariate linear models

[The material in this section is very standard and can be found in a all econometrics textbooks. For reference, you could look at: Jeffrey M. Wooldridge. 2009-2020 (depending on the edition). Introductory Econometrics: A Modern Approach. Edition 1-7. Cenage Learning. *Specifically, you could look at chapters "The simple regression model" and "Instrumental variable estimation and two-stages least squares".]*

Consider a simple linear and univariate statistical model:

$$y_i = \alpha + \beta x_i + u_i \tag{1}$$

and assume you have access to a sample of N *independent and identically distributed* observations on y_i and x_i for the population of interest. u_i represents the residual of the model, namely the set of all factors affecting the outcome y_i that are not explicitly listed as explanatory variables in the model. In our case, there is only one explanatory variable, x_i.

Our objective is the consistent estimation of the parameters of the model α and β. We proceed by imposing the following assumptions on the population statistics:

1. $E(u_i) = 0$;

2. $Cov(x_i, u_i) = 0$.

The first assumption is easily satisfied given that our model includes a constant. In most cases the constant is not of particular estimation interest and it can be interpreted as the mean of the residual. To see this, consider failure of the first assumption. For example, assume $E(u_i) = k \neq 0$. Then, we could simply rewrite the model as $y_i = (\alpha + k) + \beta x_i + (u_i - k) = y_i = \alpha_1 + \beta x_i + e_i$, with $\alpha_1 = (\alpha + k)$ and $e_i = u_i - k$. Now the residual e_i has zero mean $E(e_i) = 0$ and satisfying this condition has come at the cost of transforming the constant into $(\alpha + k)$. We will not be able to estimate α and k separately but this is usually not a major problem, unless there is a specific interest in doing so (which is usually rare. Most times we are interested in the estimation of the slope parameter β).

The second assumption, which is often called *exogeneity* of x_i, is much stronger and more subtle. $Cov(x_i, u_i) = 0$ essentially means that the explanatory variable that is explicitly listed in the model x_i is unrelated to all other factors that may matter for the determination of the outcome y_i and that are not explicitly added to the model. Of course, this is a very strong assumption.

Nevertheless, if we are willing impose these two assumptions we can then easily construct consistent estimators of the two parameters of the model α and β as follows. Let's start with the second assumption and re-write it as follows:

$$\begin{aligned} Cov(x_i, u_i) &= Cov(x_i, y_i - \alpha - \beta x_i) = Cov(x_i, y_i) - \beta Var(x_i) = 0 \\ \beta &= \frac{Cov(x_i, y_i)}{Var(x_i)} \end{aligned} \tag{2}$$

Recall that all these conditions are expressed in terms of population quantities and, hence, if we only have sample data they cannot be computed. However, with sample data they can be estimated.

Equation 2 essentially says that, if assumption $Cov(x_i, u_i) = 0$ holds, then the true value of the parameter β is equal to the ratio of the population covariance of x and y over the population variance of x. Then, if we have sample data on y and x, we can estimate this covariance and this variance to produce an estimator of β as follows:

$$\widehat{\beta} = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})(y_i - \overline{y})}{\frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})^2} \tag{3}$$

where $\overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$ is the sample average of x_i and similarly for $\overline{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$.

The expression at the numerator of equation 2 is the sample covariance of x and y and the expression at the denominator is the sample variance of x. By the *Law of Large Numbers* (and its Slutski lemma), both the numerator and the denominator converge in probability (or are consistent estimators of) the corresponding population quantities and their ratio converges in probability to the ratio of such population quantities, which, given the assumptions, is equal to the true β:

$$plim(\widehat{\beta}) = \frac{Cov(x_i, y_i)}{Var(x_i)} = \beta \tag{4}$$

To estimate the constant α we proceed similarly from the assumption $E(u_i) = 0$:

$$\begin{aligned} E(u_i) &= E(y_i - \alpha - \beta x_i) = E(y_i) - \alpha - \beta E(x_i) = 0 && (5) \\ \alpha &= E(y_i) - \beta E(x_i) && (6) \end{aligned}$$

Again, the *Law of Large Numbers* (plus its lemma) guarantees that we can obtain a consistent estimator of α by replacing the population quantities in equation 7 with consistent estimators:

$$\widehat{\alpha} = \overline{y_i} - \beta \overline{x_i} \tag{7}$$

The estimators $\{\widehat{\alpha}, \widehat{\beta}\}$ are also called the *Ordinary Least Squares* (OLS) estimators of α and β because they can also be obtained by minimising the sum of the squares residuals:

$$\{\widehat{\alpha}, \widehat{\beta}\} = argmin_{(\alpha,\beta)} \sum_{i=1}^{N} (y_i - \alpha - \beta x_i)^2 \tag{8}$$

Endogeneity and instrumental variables

Let us now see what would happen if the key assumption $Cov(x_i, u_i) = 0$ failed but we still estimated β using it. the estimator $\widehat{\beta}$ would still be a consistent estimator for the ratio of the covariance of x and y over the variance of x but, when $Cov(x_i, u_i) \neq 0$, such ratio is not equal to the true β any longer: $plim(\widehat{\beta}) = \frac{Cov(x_i,y_i)}{Var(x_i)} \neq \beta$.

To see this, consider an omitted variable z_i that is relevant for y_i (hence $\gamma \neq 0$) but for some reason we did not list it in the model. Such variable would then end up in the error term:

$$y_i = \alpha + \beta x_i + (\gamma z_i + \epsilon_i) \tag{9}$$

with $u_i = \gamma z_i + \epsilon_i$. For simplicity, let us assume that the presence of z_i is the only potential reason for failure of the assumption $Cov(x_i, u_i) = 0$, so let us assume $Cov(x_i, \epsilon_i) = 0$.

The probability limit of the estimator can be easily derived as follows:

$$\begin{aligned} plim(\widehat{\beta}) &= \frac{Cov(x_i, y_i)}{Var(x_i)} = \frac{Cov(x_i, \alpha + \beta x_i + \gamma z_i + \epsilon_i)}{Var(x_i)} \\ &= \beta + \gamma \frac{Cov(x_i, z_i)}{Var(x_i)} \end{aligned}$$

Notice that the term $\gamma\frac{Cov(x_i,z_i)}{Var(x_i)}$, which is also called the *omitted variable bias*, is zero only if either $\gamma = 0$, which means z_i is really not an omitted variable because it does not determine y_i, or $Cov(x_i, z_i) = 0$, which means that x_i and z_i are unrelated to one another. So, for the consistent estimation of β it is fine to omit variables that we can consider unrelated to the variable of interest x.

The most natural solution to the problem of omitted variables is simply not to omit them and instead include them in the explanatory variables of the model. In our example, this solution consists in estimating a model with two explanatory variables, x_i and z_i, and three parameters, α, β and γ:

$$y_i = \alpha + \beta x_i + \gamma z_i + \epsilon_i \tag{10}$$

Estimation of this model is perfectly doable but under the following assumptions:

1. $E(\epsilon_i) = 0$;
2. $Cov(x_i, \epsilon_i) = 0$;
3. $Cov(z_i, \epsilon_i) = 0$.

Unfortunately, this solution is not always feasible, for example because the omitted variables we are worried about cannot be observed in our data. In any situation in which $Cov(x_i, u_i) \neq 0$, we say that x is *endogenous* and that the model is affected by *endogeneity.*

The most popular solution to the problem of endogeneity is *instrumental variables* (IV), which is an estimation method allowing to obtain a consistent estimator of β without using the assumption $Cov(x_i, u_i) = 0$. The IV method can only be applied if we manage to find a variable, call it h_i, satisfying the following two conditions:

1. $Cov(h_i, u_i) = 0$ (exogeneity of the instrument);
2. $Cov(h_i, x_i) \neq 0$ (relevance of the instrument).

Finding such a variable may not be easy but if we find it and it is observable in our data, then we can derive the estimator using the first assumption:

$$\begin{aligned} Cov(h_i, u_i) &= Cov(h_i, y_i - \alpha - \beta x_i) = Cov(h_i, y_i) - \beta Cov(h_i, x_i) = 0 \\ \beta &= \frac{Cov(h_i, y_i)}{Cov(h_i, x_i)} \end{aligned} \tag{11}$$

As for the OLS estimator, we can construct a consistent estimator simply by replacing in the last expression the population quantities with the corresponding sample quantities:

$$\widehat{\widehat{\beta}} = \frac{\frac{1}{N}\sum_{i=1}^{N}(h_i - \overline{h})(y_i - \overline{y})}{\frac{1}{N}\sum_{i=1}^{N}(h_i - \overline{h})(x_i - \overline{x})} \tag{12}$$

By the Law of Large Numbers and its lemma, $\widehat{\widehat{\beta}}$ is a consistent estimator of β that is constructed without using the assumption $Cov(x_i, u_i) = 0$, hence it remains consistent even if such assumption may fail.