Some notes on Instrumental Variable (IV) estimation

Let's review the basic OLS estimator. Starting out with a linear form assumption:

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

note that the linear form refers to the parameters/coefficients and not to the variables.[1]

Now to obtain the OLS estimator we can use several different strategies. The most familiar one might be as the solution to the least squares problem, i.e. minimizing the sum of the squared rediuals:

$$\min_{b_0,b_1} \sum_{i=1}^{N} (y_i - b_0 + b_1 x_i)^2$$

Minimizing by finding the two first order conditions yields $b_0 = \frac{1}{N}\sum_{i=1}^{N}(y_i - b_1 x_i)$ and

$$b_1 = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2} = \frac{cov(x,y)}{var(x)}$$

Is this a good estimator? Well, if the Gauss-Markov assumptions hold the OLS estimator is BLUE (the best linear unbiased estimator). But here we will try and work with weaker assumptions (weaker in the sense: "less restrictive").

So what do we need to assume for this estimator to make sense? Well it depends on what we want our result to be. If we want to have OLS to be unbiased we need different assumptions than for consistency. Remember an unbiased estimator will get the results on average (i.e. if you draw a lot of independent random samples from the same population and take the average of the results) right, no matter the sample size. Consistency on the other hand means that as the sample size gets larger and larger we get closer and closer to the true value.

Let's start with the common assumptions: we assume the sample was randomly drawn from some population. We also assume that the regressor has a variance (i.e. that the population analogue of $var(x)$, let's call it σ_x^2, exists[2]. The difference comes in when we make assumptions about the error term. To see why let's have a closer look at what we are trying to achieve. We want to show that b_1 is a good estimator for β_1.

To do this we take the estimator b_1 and substitute the initial model for our LHS variable.

[1] It would e.g. be perfectly valid to claim that this is a linear model: $y_i = \alpha_0 + \alpha_1 \exp(w_i) + \varepsilon_i$. In fact we could simply use $x_i = \exp(w_i)$ and we are back to our original model.

[2] Exists in statistical terms means "is finite", i.e. it does not become so large as to approach infinity.

$$b_1 = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i-\bar{x})(y_i-\bar{y})}{\frac{1}{N}\sum_{i=1}^{N}(x_i-\bar{x})^2} = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i-\bar{x})(\beta_0+\beta_1 x_i+\varepsilon_i-\beta_0-\beta_1\bar{x}-\bar{\varepsilon})}{\frac{1}{N}\sum_{i=1}^{N}(x_i-\bar{x})^2}$$
$$= \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i-\bar{x})\big(\beta_1(x_i-\bar{x})+(\varepsilon_i-\bar{\varepsilon})\big)}{\frac{1}{N}\sum_{i=1}^{N}(x_i-\bar{x})^2} = \beta_1 + \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i-\bar{x})(\varepsilon_i-\bar{\varepsilon})}{\frac{1}{N}\sum_{i=1}^{N}(x_i-\bar{x})^2}$$

So the estimated coefficient and the true parameter are equal if the second part of the sum is equal to zero. Now this is not the case unless we apply some operator to the equality. If we are interested in bias, we take the expectation on both sides of the equation, when we are interested in consistency we take the probability limit instead.

Now both x and ε are random, so simply taking expectations will not get us anywhere unless we make some assumptions about the expected value of nonlinear combinations of the two. To avoid this we don't take unconditional expectations, but the expectation conditional on our regressors. This allows us to treat the regressors as "fixed".[3]

$$E(b_1|x) = \beta_1 + \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i-\bar{x})\big(E(\varepsilon_i|x)-E(\bar{\varepsilon}|x)\big)}{\frac{1}{N}\sum_{i=1}^{N}(x_i-\bar{x})^2}$$

So we need to assume that $E(\varepsilon|x)=0$ (conditional mean independence) and the OLS estimator is unbiased. Alternatively we can aim at consistency:

$$plim\, b_1 = \beta_1 + \frac{plim\frac{1}{N}\sum_{i=1}^{N}(x_i-\bar{x})(\varepsilon_i-\bar{\varepsilon})}{plim\frac{1}{N}\sum_{i=1}^{N}(x_i-\bar{x})^2} = \beta_1 + \frac{\sigma_{x\varepsilon}}{\sigma_x^2}$$

Slutsky's theorem allows us (as opposed to the situation when taking expectations) to split the problem, such that we can evaluate the limit of numerator and denominator separately.

Applying a suitable law of large numbers[4] gives us the second equality. The second sum vanishes if $\sigma_{x\varepsilon}$ is equal to zero. This is the case if $E(x\varepsilon)=0$, i.e. regressor and error are uncorrelated.

[3] $E(x|x)=x$

[4] Khinchine's weak law of large numbers if we assume that both regressors and errors are independently and identically distributed, Chebyshev's weak law of large numbers if we don't want to make the i.i.d. assumption but can assume that second moments exist and that a suitable speed of convergence applies.

Using the weaker assumption $E(x\varepsilon) = 0$ and thereby focussing on consistency, we can ask how realistic is this assumption? Sadly often it is not. One problem is that we cannot control for everything. If we think about x as the level of education a person obtains and y as (log) wages, the classical example for unobserved variables is "ability". Some individual characteristics that are related to the level of education as well as wages but ultimately unobservable. If we use w to denote "ability" we can show this easily:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 w_i + \varepsilon_i = \beta_0 + \beta_1 x_i + u_i$$

$$b_1 = \frac{cov(x,y)}{var(x)}$$

$$plim\, b_1 = \beta_1 + \beta_2 \frac{plim \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})(w_i - \bar{w})}{plim \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2} + \frac{plim \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})(\varepsilon_i - \bar{\varepsilon})}{plim \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2}$$
$$= \beta_1 + \beta_2 \frac{\sigma_{xw}}{\sigma_x^2} + \frac{\sigma_{x\varepsilon}}{\sigma_x^2}$$

Assuming $E(x\varepsilon) = 0$, i.e. the left-over part of the error term (everything that is not "ability") is uncorrelated with the level of education we have that

$$plim\, b_1 = \beta_1 + \beta_2 \frac{\sigma_{xw}}{\sigma_x^2}$$

Which is not what we want to have. OLS is inconsistent (we can also show that it is biased). Now an interesting exercise is to check what the direction (i.e. the sign) of the (asymptotic) bias is. The denominator σ_x^2 is always positive, so the sign depends on β_2 the (partial) correlation of "ability" and (log) wages and σ_{xw}, the covariance (which has the same sign as the correlation) between the level of education and "ability".

Assuming that both wages and the level of education are higher for individuals with higher "ability" means that the sign of the bias term is positive, we therefore have an upward bias (i.e. the estimated coefficient b_1 is higher than the true value β_1.

There are other reasons while the assumption on zero correlation fails, but we will save that discussion for another session.

It is important to note, that we do not have any problem with "ability" if at least one of the implicit assumptions we made does not hold. First the bias term disappears if "ability" and (log) wages are not correlated (that is if β_2 is equal to zero) or second it vanishes if "ability" and the level of education are uncorrelated (which implies that the covariance, σ_{xw}, is equal to zero).

Now if neither of that is the case, we have to find another way of dealing with the problem. The fairly simple but ingenious idea is to find some exogenous variable that induces variation in the level of education (the endogenous regressor) but is otherwise unrelated to the outcome of interest, i.e. the (log) wages. Which means that it is uncorrelated with "ability".

To make it explicit, we find an exogenous variable z, that satisfies:

$$E(zx) \neq 0 \text{ and } E(zu) = 0$$

The first assumption we can test easily, the second we have to believe[5]. So IV papers usually take great care in explaining why the instrument can be treated as exogenous.

Now if the assumption holds we can use an IV estimator or two stage least squares (2SLS) to estimate the true returns to education.

The IV estimator is simply

$$b_1^{IV} = \frac{\frac{1}{N}\sum_{i=1}^{N}(z_i - \bar{z})(y_i - \bar{y})}{\frac{1}{N}\sum_{i=1}^{N}(z_i - \bar{z})(x_i - \bar{x})} = \frac{cov(z,y)}{cov(z,x)}$$

Consistency is easy to establish by substituting the true model for y and taking probability limits. Now if we have more than one regressor or more than one instrument we cannot use simple IV but have to rely on 2SLS instead.

The idea is to use only the "good variation" in x, i.e. some part of x that is not correlated with the unobserved effects in the error term. But how do we find this "good variation"? Well we assume that the instrument is correlated with x but not with the error term u. So we can use the correlation between the instrument and the endogenous regressor. It will not be all of the "good variation" but at least some of it.

So the first stage of 2SLS is getting the correlation between our instrument the level of education. Well that is easy we can use OLS for that, running a regression of x on z (and any other regressors we might have). This will give us:

$$x_i = b_3 + b_4 z_i + v_i$$

Where v_i contains all the "bad variation" (as well as some of the "good variation") in x. So we use only the predicted value $\hat{x}_i = b_3 + b_4 z_i$. Now in the second stage we simply substitute x_i with the predicted value $\hat{x}_i$

$$b_1^{2SLS} = \frac{\frac{1}{N}\sum_{i=1}^{N}(\hat{x}_i - \bar{\hat{x}})(y_i - \bar{y})}{\frac{1}{N}\sum_{i=1}^{N}(\hat{x}_i - \bar{\hat{x}})^2}$$

To show that this estimator is consistent we plug in the true model, but use that we can decompose x into the predicted value $\hat{x}$ and a residual term v.

$$plim\ b_1^{2SLS} = \frac{\frac{1}{N}\sum_{i=1}^{N}(\hat{x}_i - \bar{\hat{x}})(\beta_0 + \beta_1\hat{x}_i + \beta_1 v_i + u_i - \beta_0 + \beta_1\bar{\hat{x}} + \beta_1\bar{v} + \bar{u})}{\frac{1}{N}\sum_{i=1}^{N}(\hat{x}_i - \bar{\hat{x}})^2} = \beta_1 + \beta_1\frac{\sigma_{\hat{x}v}}{\sigma_{\hat{x}}^2} + \frac{\sigma_{\hat{x}u}}{\sigma_{\hat{x}}^2}$$

The predicted value $\hat{x}$ is a linear function of the instrument and therefore by assumption uncorrelated with the error u (remember u combines w and ε) and by construction $\hat{x}$ and v are orthogonal so the 2SLS estimator is consistent.

[5] We can test this assumption for additional instruments, assuming that we have one valid instrument.