

# 1 Ordinary Least Squares

Last Updated: March 30, 2009

## 1.1 Introduction

We consider the linear regression model in which an outcome variable,  $y$ , is linearly related to a vector of covariates,  $(x_1, \dots, x_k)$ . This model takes the form:

$$y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik} + u_i. \quad (1)$$

This model can be written more compactly as

$$y_i = x_i \beta + u_i \quad (2)$$

where  $x_i = (1, x_1, \dots, x_k)$  and  $\beta = (\beta_0, \dots, \beta_k)'$ . We call  $u_i$  the residual term. According to some authors, if equations (1) or (2) represent an underlying *causal* relationship then we say that these equations define a **structural model**. However, others such as the Cowles Foundation have another definition in which the parameters of a structural model are policy-invariant and determine the economic behavior of firms, households and individuals. Note that with both definitions we are indeed recovering a causal relationship, but with the Cowles Foundation definition, we are recovering the actual mechanism. For the rest of these lecture notes, we will simply say that a structural model reflects a causal relationship in which a causal mechanism may or may not be specified. That said, it is important to bear in mind that understanding

mechanisms is crucial and we will hopefully discuss some of these methodological (verging on philosophical) debates towards the end of the course. If the model is not structural, it is capturing partial correlations i.e. the correlation between  $x_{i,j}$  and  $y_i$  after we control for the other covariates. Finally, I would like you to appreciate that all of the different approaches, methods and philosophies that we will discuss are simply *tools*. No tool is better than another tool. What is important is to keep an open mind, understand how to use *all* tools and have the wisdom to know which tool is best for the task at hand.

The residual plays an important role in linear regression. It contains measurement errors and omitted variables which we will discuss at greater length later on. We will maintain the following assumption on the residual term:

$$E[u_i|x_i] = 0. \tag{A1}$$

Note that because the first element of  $x_i$  is unity, this assumption implies that  $E[u_i] = 0$ . Bear in mind that this is not restrictive provided that the regression contains a constant. An alternative and stronger assumption to A1 is that  $E[u_i|x_i] = 0$ . However, this assumption is stronger than we will need. When A1 is satisfied for all the regressors in the main regression equation then we say that all of the regressors are **exogenous**.

While it is true that the conditional mean assumption is stronger than we need, it does allow us to interpret the regression equation as a conditional mean since  $E[u_i|x_i] = 0$  implies that

$$E[y_i|x_i] = \beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik}.$$

This equation implies that

$$\frac{\partial E[y_i|x_i]}{\partial x_{ij}} = \beta_j$$

so that  $\beta_j$  is the partial effect of  $x_{ij}$  on the condition mean of the dependent variable after we have controlled for the other variables in  $x_i$ .

If any of the  $k$  the orthogonality conditions in A1 fail then we will have that  $E[u_i x_{i,j}] \neq 0$  for some  $j = 1, \dots, k$ . In this case, then we say that  $x_{i,j}$  is an **endogenous** regressor. Endogenous regressors can occur for one of three reasons.

The first is omitted variables. Suppose that the true model is given by

$$w_i = \beta_0 + s_i \beta_1 + a_i \beta_2 + \varepsilon_i.$$

where  $w_i$  is the individual's wage,  $s_i$  is the individual's years of schooling and  $a_i$  is the individual's ability or intelligence. Assume that  $\varepsilon_i$  is orthogonal to all of the regressors in this equation and that  $\beta_1 > 0$  and  $\beta_2 > 0$  so that both schooling and ability result in higher wages. Because a person's ability is hard to measure, the econometrician must estimate

$$w_i = \beta_0 + s_i \beta_1 + u_i$$

where  $u_i = a_i \beta_2 + \varepsilon_i$ . In this scenario, the orthogonality conditions will fail since

$$E[u_i s_i] = E[a_i s_i] \beta_2.$$

We would expect this quantity to be positive since years of schooling and ability should be

positively related and because  $\beta_2 > 0$ .

The second cause of endogeneity is measurement error which is a problem that arises when a variable of interest is measured with some degree of inaccuracy. Suppose that the true regression equation is given by

$$y_i = \beta_0 + x_i\beta_1 + \varepsilon_i$$

where  $E[\varepsilon_i] = E[\varepsilon_i x_i] = 0$ . Next, suppose that  $x_i$  is measured with error so that we observe

$$x_i^* = x_i + e_i.$$

Assume that  $E[e_i] = E[e_i x_i] = 0$ . As a result, the econometrician must estimate

$$y_i = \beta_0 + x_i^*\beta_1 + u_i$$

where  $u_i = \varepsilon_i - \beta_1 e_i$ . The endogeneity problem results because  $Cov(u_i, x_i^*) \neq 0$ . We will discuss measurement error at greater length later on.

The third cause of endogeneity is simultaneity which is a problem that arises when the variables on both sides of regression equation are jointly determined. Consider an example in which health impacts labor supply and labor supply impacts health. The latter channel might occur either because unemployment causes stress or because unemployed people may be less likely to have access to medical care. Letting  $h_i$  and  $e_i$  denote health and employment status, this system of equations can be written as

$$h_i = \beta_0 + \beta_1 e_i + u_i$$

and

$$e_i = \alpha_0 + \alpha_1 h_i + \varepsilon_i.$$

We can solve this system and write employment status as

$$e_i = \frac{\alpha_0 + \alpha_1 \beta_0}{1 - \alpha_1 \beta_1} + \frac{\alpha_1}{1 - \alpha_1 \beta_1} u_i + \frac{1}{1 - \alpha_1 \beta_1} \varepsilon_i.$$

This equation illustrates why employment status is an endogenous variable in the health equation since it implies that

$$Cov(u_i, e_i) = \frac{\alpha_1 \sigma_u^2}{1 - \alpha_1 \beta_1} + \frac{\sigma_{u\varepsilon}}{1 - \alpha_1 \beta_1}.$$

An analogous argument can be made for estimating the impact of employment status on health.

## 1.2 Identification and Estimation

We now consider the asymptotic properties of the OLS estimator of  $\beta$  in

$$y_i = x_i \beta + u_i.$$

We assume that we observe a random sample of  $(y_i, x_i)$  of size  $n$  from some underlying population.

In addition to assumption A1, we also assume

$$rank E[x_i' x_i] = k. \tag{A2}$$

Together A1 and A2 imply that

$$E[x_i'(y - x_i\beta)] = 0 \Leftrightarrow \beta = E[x_i'x_i]^{-1} E[x_i'y_i]$$

and so these assumptions allow us to identify the parameter  $\beta$ . The analogy principle suggests that  $\beta$  can be estimated via

$$\hat{\beta} = \left( \frac{1}{n} \sum_{i=1}^n x_i'x_i \right)^{-1} \left( \frac{1}{n} \sum_{i=1}^n x_i'y_i \right).$$

This is the OLS estimator. An alternative definition of the OLS estimator that is sometimes used is

$$\hat{\beta} = \arg \min_b \sum_{i=1}^n (y_i - x_i b)^2.$$

This definition says that OLS minimizes the sum of squared residuals.

A few points deserve to be mentioned about identification. OLS will always identify  $\beta$  in equation (2) where  $E[u_i x_i'] = 0$  (provided that the rank condition is met). In fact, we can always write  $y_i$  as a linear function of  $x_i$  and an orthogonal error term. We call this function the **linear projection** of  $y_i$  onto  $x_i$ . However, the parameters of the linear projection may not correspond to the structural model. In this case, the parameters that OLS recovers are not the parameters that we care about. For example, going back to the previous example in which employment and health status are jointly determined, we observe that the structural equations are not linear projections and, thus, OLS will not recover the structural parameters.

### 1.3 Asymptotic Properties of OLS

OLS is a consistent estimator of  $\beta$ . To see this, note that we can substitute  $y_i = x_i\beta + u_i$  into the OLS formula and obtain

$$\widehat{\beta} = \beta + \left( \frac{1}{n} \sum_{i=1}^n x_i' x_i \right)^{-1} \left( \frac{1}{n} \sum_{i=1}^n x_i' u_i \right) \quad (3)$$

The Law of Large Numbers (LLN) and the continuous mapping theorem implies that the first sum will converge in probability to  $E[x_i' x_i]^{-1} \equiv A^{-1}$  and the LLN implies that the second term will converge in probability to a  $k$ -vector of zeros. By the Slutsky Theorem, we will have that  $\widehat{\beta} \xrightarrow{p} \beta$  and, thus, OLS is consistent.

Note that OLS may be biased. However, if we are willing to assume that  $E[u_i|x_i] = 0$  then we will obtain that  $E[\widehat{\beta}|x_i] = \beta$ . The Law of Iterated Expectations then gives us that  $E[\widehat{\beta}] = E[E[\widehat{\beta}|x_i]] = \beta$ .

OLS is also asymptotically normal. To see this, we re-arrange equation (3) and write

$$\begin{aligned} \sqrt{n}(\widehat{\beta} - \beta) &= \left( \frac{1}{n} \sum_{i=1}^n x_i' x_i \right)^{-1} \left( \frac{1}{\sqrt{n}} \sum_{i=1}^n x_i' u_i \right) \\ &= A^{-1} \left( \frac{1}{\sqrt{n}} \sum_{i=1}^n x_i' u_i \right) + o_p(1). \end{aligned}$$

The second term is “little o - p 1” and, thus, will converge to zero in probability. Next, the second half of the first term will converge in distribution to a normal. To see this, we note that

$E[x'_i u_i] = 0$  and apply the Central Limit Theorem to

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n x'_i u_i = \sqrt{n} \left( \frac{1}{n} \sum_{i=1}^n x'_i u_i - 0 \right) \xrightarrow{d} N(0, B)$$

where  $B \equiv E[u_i^2 x'_i x_i]$ . Putting all the pieces together, we obtain that

$$\sqrt{n} (\hat{\beta} - \beta) \xrightarrow{d} N(0, A^{-1} B A^{-1}).$$

If we further assume that

$$E[u_i^2 x'_i x_i] = \sigma^2 E[x'_i x_i] \tag{A3}$$

where  $\sigma^2 \equiv E[u_i^2]$  then assumptions A1 through A3 give us that

$$\sqrt{n} (\hat{\beta} - \beta) \xrightarrow{d} N(0, \sigma^2 A^{-1}).$$

Note that the alternative and stronger assumption that  $E[u_i^2 | x_i] = \sigma^2$  would also yield the same result. This alternative assumption tells us that the variance of the residual does not depend on the right hand side variables or that the residual is **homoskedastic**. While assumption A3 does not literally imply that the variance of the residual does not depend on  $x_i$ , we still say that it is a homoskedasticity assumption. This gives us that

$$\hat{\beta} \overset{A}{\underset{\sim}{\mathcal{L}}} N\left(\beta, \frac{\sigma^2}{n} A^{-1}\right).$$

This expression suggests that we can estimate the variance of  $\widehat{\beta}$  via

$$Avar(\widehat{\beta}) = \widehat{\sigma}^2 \left( \sum_{i=1}^n x_i' x_i \right)^{-1}$$

where  $\widehat{\sigma}^2 \equiv \frac{1}{n} \sum_{i=1}^n (y_i - x_i' \widehat{\beta})^2 \equiv \frac{1}{n} \sum_{i=1}^n \widehat{u}_i^2$ . Some people might also correct for the degrees of freedom and use  $\frac{1}{n-k} \sum_{i=1}^n \widehat{u}_i^2$ . If the sample size is large, it makes no difference whether or not we correct for degrees of freedom.

It is important to note that this asymptotic normality result is very strong. Not only does it imply that OLS is consistent, but it also implies that the OLS estimator converges to the truth (in probability) as a rate that is faster than  $\sqrt{n}$ . To see this, first note that the above result tells us that

$$\sqrt{n}(\widehat{\beta} - \beta) = O_p(1)$$

where  $O_p(1)$  means that the random variable is bounded in probability. This new concept is completely analogous to the definition of bounded for a deterministic sequence except that now we say that the probability that a sequence of random variables takes on very large values will be small. This then implies that

$$\widehat{\beta} - \beta = \frac{1}{\sqrt{n}} * O_p(1).$$

However, since the sequence  $\frac{1}{\sqrt{n}}$  will converge to zero and we know that  $o_p(1)O_p(1) = o_p(1)$ , we will have that  $\widehat{\beta} \xrightarrow{p} \beta$ . Thus, asymptotic normality implies consistency. Next, we note that  $\frac{n^c}{n^{1/2}}$

will converge to zero for  $0 \leq c < \frac{1}{2}$  and, thus, we will have that

$$p \lim n^c (\widehat{\beta} - \beta) = 0 \text{ for } 0 \leq c < \frac{1}{2}.$$

Consequently, we can conclude that OLS will converge to the truth faster than  $n^c$  goes to infinity.

## 1.4 Heteroskedastic Robust Inference

Suppose that A3 does not hold. Then we say that the residual is **heteroskedastic** or that **heteroskedasticity** is present. Then the asymptotic variance will be given by  $A^{-1}BA^{-1}$  which we estimate by

$$Avar(\widehat{\beta}) = \left( \sum_{i=1}^n x_i' x_i \right)^{-1} \left( \sum_{i=1}^n \widehat{u}_i^2 x_i' x_i \right) \left( \sum_{i=1}^n x_i' x_i \right)^{-1}.$$

These are the **Eicker-White Standard Errors** or **Robust Standard Errors**. In practice, researchers almost always use the robust standard errors when using OLS. Also, note that heteroskedasticity does not affect the consistency or asymptotic normality of the OLS estimator, it just affects its efficiency *i.e.* the asymptotic variance. However, when estimating non-linear regression models such as the Probit and, particularly, the Tobit model, heteroskedasticity *will* affect consistency and, thus, have a potentially more pernicious impact. In general, there are not many convincing methods of dealing with heteroskedasticity in non-linear models. In practice, heteroskedasticity does not seem to matter too much in the Probit model, but it can have quite devastating effects in the Tobit model.

An alternative means of dealing with heteroskedasticity is **Weighted Least Squares**. Suppose that  $E[u_i^2|x_i] = \sigma_{x_i}^2$  and that we know the functional form. First, we estimate  $\widehat{\sigma}_{x_i}^2$ . Second,

we transform the data to

$$\tilde{x}_i = \frac{x_i}{\hat{\sigma}_{x_i}}; \tilde{y}_i = \frac{y_i}{\hat{\sigma}_{x_i}}.$$

Third, we run OLS on the transformed data and calculate the weighted least squares estimator:

$$\hat{\beta}^{WLS} = \left( \frac{1}{n} \sum_{i=1}^n \tilde{x}_i' \tilde{x}_i \right)^{-1} \left( \frac{1}{n} \sum_{i=1}^n \tilde{x}_i' \tilde{y}_i \right).$$

We will then have that

$$\sqrt{n} \left( \hat{\beta}^{WLS} - \beta \right) = E [\tilde{x}_i' \tilde{x}_i]^{-1} \left( \frac{1}{\sqrt{n}} \sum_{i=1}^n \tilde{x}_i' \tilde{u}_i \right) + o_p(1) \xrightarrow{d} N \left( 0, E [\tilde{x}_i' \tilde{x}_i]^{-1} \right)$$

provided that  $E[u|x_i] = 0$ . Note that we require the conditional mean assumption and not simply the assumption that the residual and the regressors are uncorrelated. To better understand, why we obtained this asymptotic variance note that  $E[\tilde{u}_i^2 \tilde{x}_i' \tilde{x}_i] = E\left[\frac{u_i^2}{\sigma_x^2} \tilde{x}_i' \tilde{x}_i\right] = E\left[E\left[\frac{u_i^2}{\sigma_x^2} \tilde{x}_i' \tilde{x}_i | x_i\right]\right] = E[\tilde{x}_i' \tilde{x}_i]$ . This estimator is more efficient than OLS. However, it requires knowledge of the functional form for  $\sigma_{x_i}^2$ .

## 1.5 Frisch-Waugh Theorem

Consider the linear projection of  $y_i$  onto  $x_i$  and  $z_i$  :

$$y_i = x_i \beta + z_i \gamma + u_i$$

where  $E[u_i x_i'] = E[u_i z_i'] = 0$ . Alternatively, we can write

$$L(y_i|z_i) = L(x_i|z_i)\beta + z_i\gamma$$

where  $L(a|b)$  denotes the linear projection of  $a$  onto  $b$ . Thus, we will have that

$$\underbrace{y_i - L(y_i|z_i)}_{r_{y_i}} = \underbrace{(x_i - L(x_i|z_i))}_{r_{x_i}}\beta + u_i$$

and we can conclude that

$$\beta = E[r'_{x_i} r_{x_i}]^{-1} E[r'_{x_i} r_{y_i}].$$

This result tells us that we can identify  $\beta$  in the long regression if we regress the residuals from the short regression of  $y_i$  onto  $z_i$  on the residuals from the short regression of  $x_i$  onto  $z_i$ .

## 1.6 Measurement Error

Suppose that the population model is given by

$$y_i = \beta_0 + \beta_1 x_{i,1} + \dots + \beta_k x_{i,k} + u_i$$

where  $E[u_i x_{i,j}] = 0$  for  $j = 1, \dots, k$ . However, the econometrician does not observe  $x_{i,k}$ . Rather, she observes  $x_{i,k}$  with error:

$$x_{i,k}^* = x_{i,k} + e_i$$

where  $E[e_i] = E[e_i x_{i,j}] = 0$  for  $j = 1, \dots, k$ . In addition, we will assume that  $E[e_i u_i] = 0$ . In practice, measurement errors result from the inability of surveys to perfectly measure variables. For example, income and consumption expenditures are notorious for being mis-measured. Consequently, the econometrician must estimate the parameters in the regression:

$$y_i = \beta_0 + \beta_1 x_{i,1} + \dots + \beta_k x_{i,k}^* + (u_i - \beta_k e_i).$$

This is problematic since the residual term will be correlated with  $x_{i,k}^*$  which implies that OLS will not yield consistent estimates.

We will now calculate the asymptotic bias in the OLS estimate of  $\beta$ . First, we will project  $x_{i,k}$  onto  $(1, x_{i,1}, \dots, x_{i,k-1})$  and obtain

$$x_{i,k} = \delta_0 + \delta_1 x_{i,1} + \dots + \delta_{k-1} x_{i,k-1} + r_{i,k}$$

which implies that

$$x_{i,k}^* = \delta_0 + \delta_1 x_{i,1} + \dots + \delta_{k-1} x_{i,k-1} + r_{i,k} + e_k$$

where  $E[r_{i,k} x_{i,j}] = 0$  for  $j = 1, \dots, k-1$ . Next, note that OLS will recover the parameters from the following linear projection:

$$y_i = \tilde{\beta}_0 + \tilde{\beta}_1 x_{i,1} + \dots + \tilde{\beta}_k x_{i,k}^* + \tilde{u}_i$$

where the residual  $\tilde{u}_i$  is orthogonal to all of the regressors. The Frisch-Waugh Theorem implies

that

$$\tilde{\beta}_k = E [(r_{i,k} + e_k)^2]^{-1} E [(r_{i,k} + e_k)y_i] = \beta_k \frac{\sigma_{r_k}^2}{\sigma_{r_k}^2 + \sigma_e^2}.$$

Note that because  $\frac{\sigma_{r_k}^2}{\sigma_{r_k}^2 + \sigma_e^2} < 1$ , we will have that  $|\tilde{\beta}_k| < |\beta_k|$ . Thus, classical measurement error will cause OLS to be biased towards zero.

Also, note that if we only have one regressor then  $\sigma_r^2$  will be the variance of the residual from a regression of  $x_i$  onto a constant i.e.

$$x_i = \mu + r_i$$

where  $\mu = E[x_i]$ . Thus, we will have that  $\sigma_r^2 = \sigma_x^2$  in which case the bias formula simplifies to

$$\tilde{\beta} = \beta \frac{\sigma_x^2}{\sigma_x^2 + \sigma_e^2}.$$

Interestingly, in the case where we only have one regressor, the reverse regression of  $x_i^*$  onto  $y_i$  and a constant will identify the parameter

$$\beta^R = \frac{Cov(x_i^*, y_i)}{Var(y_i)} = \frac{\beta \sigma_x^2}{\beta^2 \sigma_x^2 + \sigma_u^2}$$

and so, we will have that

$$\frac{1}{\beta^R} = \beta + \frac{\sigma_u^2}{\beta \sigma_x^2}.$$

This is an interesting result because it implies that if  $\beta > 0$  then we will have that

$$\beta \in \left[ \tilde{\beta}, \frac{1}{\beta^R} \right]$$

and if  $\beta < 0$  then we will have that

$$\beta \in \left[ \frac{1}{\beta^R}, \tilde{\beta} \right].$$

Measurement error in the dependent variable typically poses less of a problem than measurement error in the independent variable. To see this, suppose that the econometrician observes

$$y_i^* = y_i + e_i$$

and estimates

$$y_i^* = \beta_0 + \beta_1 x_i + u_i \Leftrightarrow y_i = \beta_0 + \beta_1 x_i + u_i - e_i.$$

Clearly, as long as the measurement error is orthogonal to the right-hand side regressors then OLS will still yield consistent estimates. The only cost of measurement error in this case will be an efficiency loss.

## 1.7 Goodness-of-Fit

We now discuss how we can measure how well the OLS regression fits the data. In the linear regression model of equation (1), we will have that

$$\sigma_y^2 = \sigma_{x\beta}^2 + \sigma_u^2.$$

We define

$$\rho^2 = \frac{\sigma_{x\beta}^2}{\sigma_{x\beta}^2 + \sigma_u^2} = 1 - \frac{\sigma_u^2}{\sigma_y^2}.$$

This object, which we call the **population**  $R^2$ , tells us that percentage of variation in the dependent variable that is explained by the independent variables. In the sample, we will have

$$R^2 = 1 - \frac{SSR}{TSS}$$

where  $SSR \equiv \sum_{i=1}^n \hat{u}_i^2$  and  $TSS \equiv \sum_{i=1}^n (y_i - \bar{y})^2$ . We call  $SSR$ , the sum of squared residuals, and,  $TSS$ , the total sum of squares.

## 1.8 More on Omitted Variables

We conclude with a more in-depth discussion of omitted variables. Suppose that the true model is given by

$$y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik} + \gamma q_i + v_i$$

where  $E[v_i x_{ij}] = 0$  for  $j = 1, \dots, k$ . However, the econometrician does not observe  $q_i$  and is, thus, forced to estimate

$$y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik} + u_i$$

where  $u_i = \gamma q_i + v_i$ . We can write

$$q_i = \delta_0 + \delta_1 x_{i1} + \dots + \delta_k x_{ik} + r_i$$

where  $E[r_i x_{ij}] = 0$  for  $j = 1, \dots, k$ . This then gives us that

$$y_i = (\beta_0 + \gamma \delta_0) + (\beta_1 + \gamma \delta_1) x_{i1} + \dots + (\beta_k + \gamma \delta_k) x_{ik} + (u_i + \gamma r_i)$$

and we can conclude that

$$p \lim \widehat{\beta}_j = \beta_j + \gamma \delta_j$$

for  $j = 1, \dots, k$ . In order for omitted variables to be an issue, we must have that both  $\gamma \neq 0$  and that  $\delta_j \neq 0$ .