Linear models

Linear models are computationally convenient and remain widely used in applied econometric research

Our main focus in these lectures will be on single equation linear models of the form

$$y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_K x_{Ki} + u_i$$

where we have:

- a scalar outcome variable y_i for observation i
- a set of K explanatory variables $x_{1i}, x_{2i}, ..., x_{Ki}$

and a sample of n observations indexed $i = 1, 2, ..., n$

$$y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_K x_{Ki} + u_i$$

The linear specification is not quite as restrictive as it may seem at first sight

The dependent variable and each of the explanatory variables may be obtained as known, non-linear transformations of other underlying variables

For example, we may have $y_i = \ln Y_i$, or we may have $x_{3i} = x_{2i}^2$

The restrictive features of the linear model are that the parameters β_1, β_2,... ...,β_K enter linearly, and the error term u_i enters additively

$$y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_K x_{Ki} + u_i$$

Linear models will usually have an unrestricted intercept term

For example, with $K = 2$, we may have $x_{1i} = 1$ for all $i = 1, 2, ...n$, in which case the parameter β_1 corresponds to the intercept of a straight line

With $K = 3$, we may have $x_{1i} = 1$ for all $i = 1, 2, ...n$, in which case the parameter β_1 corresponds to the intercept of a plane

More generally, if $x_{1i} = 1$ for all $i = 1, 2, ...n$, the parameter β_1 corresponds to the intercept of a K-dimensional hyperplane

The error term u_i reflects the combined effect of all additional influences on the outcome variable y_i

This may include the effect of relevant variables not included in our set of K explanatory variables, and also any non-linear effects of our included variables $(x_{1i}, x_{2i}, ..., x_{Ki})$

Notation

$$y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_K x_{Ki} + u_i$$

Letting x_i and β denote the $K \times 1$ column vectors

$$x_i = \begin{pmatrix} x_{1i} \\ x_{2i} \\ \vdots \\ x_{Ki} \end{pmatrix} \quad \text{and} \quad \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_K \end{pmatrix}$$

we can write the model as

$$y_i = x_i'\beta + u_i \quad \text{for } i = 1, ..., n$$

Sometimes you may see the i subscripts omitted and the model written as

$$y = \beta_1 x_1 + \beta_2 x_2 + ... + \beta_K x_K + u$$

or

$$y = x'\beta + u$$

for a typical observation (population model)

We can also stack the observations for a sample of size n, using

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad X = \begin{pmatrix} x_1' \\ x_2' \\ \vdots \\ x_n' \end{pmatrix} = \begin{pmatrix} x_{11} & x_{21} & \cdots & x_{K1} \\ x_{12} & x_{22} & \cdots & x_{K2} \\ \vdots & \vdots & \vdots & \vdots \\ x_{1n} & x_{2n} & \vdots & x_{Kn} \end{pmatrix}, \quad u = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix}$$

where y and u are $n \times 1$ column vectors and X is an $n \times K$ matrix, to obtain

$$y = X\beta + u$$

NB. It should be clear from the context whether y and u denote $n \times 1$ vectors as here, or scalars for a typical observation, as on the previous slide

Ordinary least squares (OLS)

The OLS estimator is a widely used method for estimating the unknown parameter vector β from the data on y_i and $x_{1i}, ..., x_{Ki}$ for a sample of $i = 1, ..., n$ observations

The OLS estimator finds the parameter vector β which minimizes the sum of the squared errors $u_i = y_i - x_i'\beta = u_i(\beta)$

Formally

$$\begin{aligned}\widehat{\beta}_{\text{OLS}} &= \arg\min_{\beta} \sum_{i=1}^{n} u_i^2(\beta) = \arg\min_{\beta} \sum_{i=1}^{n} (y_i - x_i'\beta)^2 \\ &= \arg\min_{\beta} \quad u(\beta)'u(\beta) = \arg\min_{\beta} (y - X\beta)'(y - X\beta)\end{aligned}$$

This squared error loss function penalizes large values of the error term, but tolerates small values of the error term

The OLS estimator finds a value of the parameter vector β which avoids large residuals, while tolerating many small deviations from the fitted model

This estimator can be shown to have several desirable properties in linear models of this kind

One drawback is that the OLS parameter estimates can be sensitive to the inclusion or exclusion of 'outliers' (influential observations) from the sample, and it is usually good practice to check the robustness of the results to changes in the sample

The OLS estimator

To obtain the OLS estimator, we minimize $\phi(\beta) = \sum_{i=1}^{n} u_i^2(\beta)$ with respect to β, where $u_i(\beta) = y_i - \beta_1 x_{1i} - ... - \beta_K x_{Ki}$

Consider

$$\frac{\partial \phi(\beta)}{\partial \beta_k} = \sum_{i=1}^{n} 2u_i(\beta) \left(\frac{\partial u_i(\beta)}{\partial \beta_k} \right) = \sum_{i=1}^{n} -2u_i(\beta) x_{ki} \quad \text{for } k = 1, ..., K$$

Setting

$$\frac{\partial \phi(\beta)}{\partial \beta_k} = 0 \leftrightarrow \sum_{i=1}^{n} x_{ki} u_i(\beta) = 0 \quad \text{for } k = 1, ..., K$$

These K first-order conditions can be written more compactly as $X'u(\beta) = 0$, since

$$X'u(\beta) = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ \vdots & \vdots & \vdots & \vdots \\ x_{K1} & x_{K2} & \cdots & x_{Kn} \end{pmatrix} \begin{pmatrix} u_1(\beta) \\ u_2(\beta) \\ \vdots \\ u_n(\beta) \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{n} x_{1i}u_i(\beta) \\ \vdots \\ \sum_{i=1}^{n} x_{Ki}u_i(\beta) \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}$$

where X' is $K \times n$, $u(\beta)$ is $n \times 1$, and so $X'u(\beta)$ is $K \times 1$

The first-order conditions thus set $X'u(\beta) = 0$

Now using $u(\beta) = y - X\beta$, we obtain

$$X'(y - X\beta) = 0$$

$$\leftrightarrow X'y - X'X\beta = 0$$

or a system of K 'normal equations' for the K unknown elements of β

$$X'y = X'X\beta$$

Now provided $(X'X)^{-1}$ exists, the parameter vector β which minimizes $\phi(\beta) = \sum_{i=1}^{n} u_i^2(\beta)$ satisfies

$$\beta = (X'X)^{-1}X'y$$

This gives the OLS estimator as

$$\begin{aligned} \widehat{\beta}_{\text{OLS}} &= (X'X)^{-1}X'y \\ &= \left(\sum_{i=1}^{n} x_i x_i'\right)^{-1} \left(\sum_{i=1}^{n} x_i y_i\right) \end{aligned}$$

recalling that $y_i = x_i'\beta + u_i$ and x_i is a $K \times 1$ vector

The OLS estimator can also be written as

$$\widehat{\beta}_{\text{OLS}} = \left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} x_i y_i\right)$$

In the simple case of an intercept and a single explanatory variable, i.e. $y_i = \beta_1 + \beta_2 x_{2i} + u_i$, this expression implies that

$$\widehat{\beta}_1 = \overline{y} - \widehat{\beta}_2 \overline{x}_2$$

$$\widehat{\beta}_2 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_{2i} - \overline{x}_2)(y_i - \overline{y})}{\frac{1}{n}\sum_{i=1}^{n}(x_{2i} - \overline{x}_2)^2}$$

where $\widehat{\beta}_1$ and $\widehat{\beta}_2$ are the OLS estimates of the intercept and slope parameters, and $\overline{y}$ and $\overline{x}_2$ are the sample means of y_i and x_{2i}

$$\widehat{\beta}_1 = \overline{y} - \widehat{\beta}_2 \overline{x}_2 \leftrightarrow \overline{y} = \widehat{\beta}_1 + \widehat{\beta}_2 \overline{x}_2$$

The fitted relationship using OLS passes through the sample means (this result generalizes to models with more than one explanatory variable)

$$\widehat{\beta}_2 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_{2i} - \overline{x}_2)(y_i - \overline{y})}{\frac{1}{n}\sum_{i=1}^{n}(x_{2i} - \overline{x}_2)^2}$$

The estimated slope parameter is the ratio of the sample covariance between x_{2i} and y_i and the sample variance of x_{2i} (this result does not generalize to models with more than one explanatory variable)

$$\widehat{\beta}_1 = \overline{y} - \widehat{\beta}_2 \overline{x}_2$$

$$\widehat{\beta}_2 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_{2i} - \overline{x}_2)(y_i - \overline{y})}{\frac{1}{n}\sum_{i=1}^{n}(x_{2i} - \overline{x}_2)^2}$$

Replacing y_i by $y_i^* = y_i - \overline{y}$ and replacing x_{2i} by $x_{2i}^* = x_{2i} - \overline{x}_2$ would leave the OLS estimate of the slope parameter unchanged, and make the OLS estimate of the intercept zero

This result does generalize to models with more than one explanatory variable

We can obtain the same OLS estimates of the slope parameters either from the model

$$y_i = \beta_1 + \beta_2 x_{2i} + ... + \beta_K x_{Ki} + u_i$$

or from the model expressed in terms of deviations from sample means

$$y_i^* = \beta_2 x_{2i}^* + ... + \beta_K x_{Ki}^* + u_i^*$$

Unless stated otherwise, all the linear models we consider will either contain an intercept term, or be expressed in terms of deviations of original variables from their sample means

In the special case with $K = 1$ and $x_{1i} = 1$ for $i = 1, 2, ..., n$, i.e. the model $y_i = \beta_1 + u_i$, the first-order condition $\sum_{i=1}^{n} u_i(\beta_1) = 0$ implies that

$$\sum_{i=1}^{n}(y_i - \beta_1) = \left(\sum_{i=1}^{n} y_i\right) - n\beta_1 = 0 \leftrightarrow \left(\sum_{i=1}^{n} y_i\right) = n\beta_1$$

so that the OLS estimator $\widehat{\beta}_1 = \frac{1}{n}\sum_{i=1}^{n} y_i = \overline{y}$ is the sample mean of y_i

We should check that, by finding the solution to the first-order conditions, we have located a minimum and not a maximum of $\phi(\beta) = u(\beta)'u(\beta)$, where $u(\beta) = y - X\beta$

$\widehat{\beta}_{\text{OLS}}$ solves the first-order conditions, so that $X'\widehat{u} = 0$, where $\widehat{u} = y - X\widehat{\beta}_{\text{OLS}}$

For any candidate value of the parameter vector, say $\widetilde{\beta}$, let $d = \widetilde{\beta} - \widehat{\beta}_{\text{OLS}}$, or equivalently write $\widetilde{\beta} = \widehat{\beta}_{\text{OLS}} + d$

Then $\widetilde{u} = y - X\widetilde{\beta} = y - X(\widehat{\beta}_{\text{OLS}} + d) = y - X\widehat{\beta}_{\text{OLS}} - Xd = \widehat{u} - Xd$

And so $\widetilde{u}'\widetilde{u} = (\widehat{u} - Xd)'(\widehat{u} - Xd)$

Now since $X'\widehat{u} = 0$, and hence $\widehat{u}'X = 0$, we have $\phi(\widetilde{\beta}) =$

$$\begin{aligned}
\widetilde{u}'\widetilde{u} &= (\widehat{u} - Xd)'(\widehat{u} - Xd) \\
&= \widehat{u}'\widehat{u} - \widehat{u}'Xd - d'X'\widehat{u} + d'X'Xd \\
&= \widehat{u}'\widehat{u} + d'X'Xd \\
&= \widehat{u}'\widehat{u} + v'v
\end{aligned}$$

where $v = Xd$ is an $n \times 1$ column vector

Since $v'v$ is a scalar sum of squares, we know that $v'v \geqslant 0$, with $v'v = 0$ if and only if $v = 0$

Thus $\phi(\widetilde{\beta}) \geqslant \phi(\widehat{\beta}_{\mathrm{OLS}}) = \widehat{u}'\widehat{u}$, with $\phi(\widetilde{\beta}) = \phi(\widehat{\beta}_{\mathrm{OLS}})$ iff $v = Xd = 0$

The OLS estimator $\widehat{\beta}_{\text{OLS}}$ does indeed minimize $\phi(\beta) = u(\beta)'u(\beta)$, as desired

Moreover if $\text{rank}(X) = K$, the only $K \times 1$ vector d that satisfies $Xd = 0$ is $d = 0$, so that $\phi(\widetilde{\beta}) = \phi(\widehat{\beta}_{\text{OLS}})$ iff $\widetilde{\beta} = \widehat{\beta}_{\text{OLS}}$

In this case, the solution to the first-order conditions gives us the unique minimum

Note that if $\text{rank}(X) < K$, then $(X'X)^{-1}$ does not exist, and there is not a unique solution to the normal equations for all K elements of the parameter vector β

This situation, known as 'perfect multicollinearity', is easily avoided in practice by not including any explanatory variables that are perfect linear combinations of a set of other explanatory variables included in the model

For example, if the model includes x_{2i} and x_{3i}, we could not also include a variable $x_{4i} = x_{2i} - x_{3i}$, or $x_{5i} = x_{2i} + x_{3i}$

If you are familiar with matrix calculus, you can verify that minimizing $\phi(\beta) = u(\beta)'u(\beta) = (y - X\beta)'(y - X\beta)$ with respect to the parameter vector β yields the first-order conditions

$$\frac{\partial \phi(\beta)}{\partial \beta} = -2X'(y - X\beta) = 0$$

and the second-order conditions

$$\frac{\partial^2 \phi(\beta)}{\partial \beta \partial \beta'} = 2X'X > 0$$

if the $K \times K$ matrix $X'X$ is positive definite, confirming that we obtain a unique minimum if $(X'X)^{-1}$ exists

Regression

The general regression model with additive errors is written in matrix notation as

$$y = \mathrm{E}(y|X) + u$$

This is simply an additive decomposition of the random vector y into its conditional expectation given X, and a vector of additive deviations from this conditional expectation

To motivate an initial focus on regression models, we can note some properties of the conditional expectation function

Suppose y_i is a random variable, x_i is a $K \times 1$ random vector, and we are interested in *predicting* the value taken by y_i using only information on the values taken by the random variables in x_i; that is, we consider predictions of the form $h(x_i)$, where $h(x_i)$ is a function which maps from the K dimensional vector x_i to scalars

The particular function $h^*(x_i)$ that minimises the mean squared error

$$\text{MSE}(h) = \text{E}[y_i - h(x_i)]^2$$

is *unique*, and given by the *conditional expectation function* $h^*(x_i) = \text{E}(y_i|x_i)$

NB. Notation - Previously, for 2 random variables Y and X, we wrote the conditional expectation of Y given some value taken by X as $\mathrm{E}(Y|X = x)$, and noted that this can be viewed as a function of x

The conditional expectation function is often denoted by $\mathrm{E}(Y|X)$

Evaluated at a particular value of X, this function returns the conditional expectation of Y given that X takes that particular value

For 2 random variables y_i and x_i, this becomes $\mathrm{E}(y_i|x_i)$

And in the case where x_i is a random vector, this function evaluated at a particular value of x_i returns the conditional expectation of y_i given that each element of the random vector x_i takes those particular values

The conditional expectation function is also the unique function of x_i with the orthogonality property that, for every other bounded function $g(x_i)$:

$$\mathrm{E}\left\{[y_i - h^*(x_i)]g(x_i)\right\} = 0$$

Some properties of the conditional expectation function which follow from this orthogonality condition include:

i) If $x_i = 1$ (or any other constant) then $\mathrm{E}(y_i|x_i) = \mathrm{E}(y_i)$

ii) If y_i and (all the elements of) x_i are independent, then $\mathrm{E}(y_i|x_i) = \mathrm{E}(y_i)$

iii) If another random vector $z_i \perp\!\!\!\perp (y_i, x_i)$, then $\mathrm{E}(y_i|x_i, z_i) = \mathrm{E}(y_i|x_i)$

iv) If y_{1i} and y_{2i} are random variables and a_1 and a_2 are constants, then $\mathrm{E}[(a_1 y_{1i} + a_2 y_{2i})|x_i] = a_1 \mathrm{E}(y_{1i}|x_i) + a_2 \mathrm{E}(y_{2i}|x_i)$

Linear Regression

A linear regression model is obtained if the conditional expectation of y given X is a linear function of X, i.e. $\mathrm{E}(y|X) = X\beta$

This gives

$$y = X\beta + u$$

$$\text{or} \quad y_i = x_i'\beta + u_i \quad \text{for } i = 1, ..., n$$

with $\mathrm{E}(u_i|x_i) = 0$ (scalar)

and $\mathrm{E}(u|X) = 0$ ($n \times 1$ vector)

One motivation for the OLS estimator is that it will be shown to have some desirable properties as an estimator of the parameter vector β in a class of linear regression models

One context in which the conditional expectation of y_i given x_i would be a linear function of x_i is the case in which the random vector $(y_i, x_i')'$ has a normal distribution, i.e. when all the variables considered in the model are jointly normal

Suppose we have $K = 2$, $x_{1i} = 1$, and the random vector

$$\begin{pmatrix} y_i \\ x_{2i} \end{pmatrix} \sim \mathrm{N}\left[\begin{pmatrix} \mu_y \\ \mu_{x_2} \end{pmatrix}, \begin{pmatrix} \sigma_y^2 & \sigma_{yx_2} \\ \sigma_{yx_2} & \sigma_{x_2}^2 \end{pmatrix} \right]$$

Then the *marginal distribution* $y_i \sim \mathrm{N}(\mu_y, \sigma_y^2)$ is also normal

And the *conditional distribution* $y_i|x_{2i} \sim \mathrm{N}(\beta_1 + \beta_2 x_{2i}, \sigma_y^2 - \beta_2^2 \sigma_{x_2}^2)$ where $\beta_1 = \mu_y - \beta_2 \mu_{x_2}$ and $\beta_2 = \sigma_{yx_2}/\sigma_{x_2}^2$

The conditional distribution is also normal, with the *conditional expectation* $\mathrm{E}(y_i|x_{2i}) = \mathrm{E}(y_i|x_i) = \beta_1 + \beta_2 x_{2i}$, a linear function of x_i

NB. Compare these expressions for β_1 and β_2 with the expressions we had for the OLS estimators $\widehat{\beta}_1$ and $\widehat{\beta}_2$ in the case with $K = 2$

Sample means replace expected values, and sample (co-)variances replace (co-)variances

More generally, in empirical work in economics, we may have little reason to believe the conditional expectation of y given X is the linear function $X\beta$

The OLS estimator is also commonly used in models where we may not believe the linear conditional expectation specification

We may still be interested in the **linear projection** of y on X, denoted $\mathrm{E}^*(y|X) = X\beta$

The predicted values of $\widehat{y} = X\widehat{\beta}_{\mathrm{OLS}}$ still have an optimality property, in the sense of minimising the mean squared error within the class of predicted values that can be constructed as linear functions of X

Linear models for which we do not take seriously the linear specification of the conditional expectation function are not, strictly speaking, *linear regression* models, although this term is often used more generally in applied econometrics

The OLS estimates of the parameters in linear models may still provide a useful summary of patterns (partial correlations) in the data, although we should be extremely cautious about attaching any causal significance to the estimated parameters (since correlation does not imply causation)

This is the case *whether or not* we consider the linear function $X\beta$ to be the conditional expectation function $\mathrm{E}(y|X)$

Example:

Suppose the process which generates the outcomes y_i is the linear process

$$y_i = \gamma_1 + \gamma_2 x_{2i} + \gamma_3 x_{3i} + u_i$$

with $u_i \perp\!\!\!\perp (x_{2i}, x_{3i})$ so that $\mathrm{E}(u_i|x_{2i}, x_{3i}) = 0$ and $\mathrm{E}(u_i|x_{2i}) = \mathrm{E}(u_i|x_{3i}) = 0$

Then we have

$$\mathrm{E}(y_i|x_{2i}, x_{3i}) = \gamma_1 + \gamma_2 x_{2i} + \gamma_3 x_{3i}$$

Suppose the process which generates the explanatory variable x_{3i} is the linear process

$$x_{3i} = \delta_1 + \delta_2 x_{2i} + v_i$$

with $v_i \perp\!\!\!\perp x_{2i}$ so that $\mathrm{E}(v_i|x_{2i}) = 0$

We also have

$$\mathrm{E}(x_{3i}|x_{2i}) = \delta_1 + \delta_2 x_{2i}$$

Now suppose we have data on y_i and x_{2i}, but not on x_{3i}

We still have a linear conditional expectation function

$$\begin{aligned}
\mathrm{E}(y_i|x_{2i}) &= \gamma_1 + \gamma_2 x_{2i} + \gamma_3 \mathrm{E}(x_{3i}|x_{2i}) \\
&= \gamma_1 + \gamma_2 x_{2i} + \gamma_3(\delta_1 + \delta_2 x_{2i}) \\
&= (\gamma_1 + \gamma_3\delta_1) + (\gamma_2 + \gamma_3\delta_2)x_{2i} \\
&= \beta_1 + \beta_2 x_{2i}
\end{aligned}$$

If our interest is in *predicting* the value taken by y_i using only information on the value taken by x_{2i}, we will be interested in estimating the parameters of this linear conditional expectation function

If however we are interested in learning about the *ceteris paribus* effect of a change in x_{2i} on the outcome y_i, holding x_{3i} constant, this is given by the parameter γ_2 in the process which generates y_i, and we have $\beta_2 = \gamma_2 + \gamma_3\delta_2 \neq \gamma_2$ (except in special cases)

In this case, we will need to do something different to learn about γ_2; we will return to this problem later in the course

[Example from economics: y_i is a measure of wages, x_{2i} is a measure of educational attainment, x_{3i} is an unmeasured characteristic such as ability]

For now, notice that linearity of the conditional expectation function $\mathrm{E}(y_i|x_{2i})$ does not endow the OLS estimates of the parameters of this linear function with any causal significance

Classical linear regression models

To establish properties of the OLS estimator in particular versions of the linear model, we need to make some more specific assumptions about the model

We start by considering assumptions under which we can derive properties that hold in small samples (exact finite sample properties), as well as in large samples

Specific versions of the linear model that allow us to derive these exact finite sample properties are sometimes referred to as 'classical linear regression models'

Later we will see that the OLS estimator may still have some desirable properties in large samples (asymptotic properties) under weaker assumptions

The classical linear regression model with fixed regressors

Some treatments start by assuming that the error term (u_i) and the outcome variable (y_i) are random variables, or are 'stochastic', but that the explanatory variables ($x_{i1}, ..., x_{Ki}$) are not random variables, or are 'non-stochastic'

The explanatory variables are then described as being 'fixed in repeated samples', or 'fixed'

This reflects the development of statistical methods for linear regression models to analyze the outcomes of controlled experiments (e.g. drug trials), in which the level of the dose is controlled by the researcher

This assumption is rarely appropriate for the kinds of data used in applied econometrics, which are mostly not generated by controlled experiments (except perhaps when considering laboratory data in experimental economics, or randomised control trials in some contexts)

For example, a labour economist may be interested in modelling labour supply as a function of wages, but may also be interested in modelling the determinants of wages

In the first case the wage variable is an explanatory variable, in the second case the wage variable is the outcome of interest

Similarly a macroeconomist may be interested in the effects of exchange rates on imports and exports, but may also be interested in modelling exchange rates

It makes no sense to assume that the data on wages or exchange rates are stochastic when these variables appear as dependent variables, but to treat them as non-stochastic when they appear as explanatory variables in other models

The classical linear regression model with fixed regressors can be stated as

$$y = X\beta + u$$

$$\mathrm{E}(u) = 0, \quad \mathrm{Var}(u) = \sigma^2 I$$

$$X \text{ is non-stochastic and full rank}$$

or equivalently as

$$\mathrm{E}(y) = X\beta$$

$$\mathrm{Var}(y) = \sigma^2 I$$

$$X \text{ is non-stochastic and full rank}$$

Notice that $\mathrm{Var}(u) = \sigma^2 I$ implies $\mathrm{Cov}(u_i, u_j) = 0$ for all $i \neq j$

Under these assumptions we can show that

$$\mathrm{E}(\widehat{\beta}_{\mathrm{OLS}}) = \beta$$

The OLS estimator $\widehat{\beta}_{\mathrm{OLS}}$ is said to be an **unbiased** estimator of the parameter vector β

$$\mathrm{Var}(\widehat{\beta}_{\mathrm{OLS}}) = \sigma^2 (X'X)^{-1}$$

$$\mathrm{Var}(\widehat{\beta}_{\mathrm{OLS}}) \leqslant \mathrm{Var}(\widetilde{\beta})$$

for any other unbiased estimator $\widetilde{\beta}$ that is a linear function of the random vector y

NB. For $K > 1$, this means that the $K \times K$ matrix $\mathrm{Var}(\widetilde{\beta}) - \mathrm{Var}(\widehat{\beta}_{\mathrm{OLS}}) \geqslant 0$ (i.e. $\mathrm{Var}(\widetilde{\beta}) - \mathrm{Var}(\widehat{\beta}_{\mathrm{OLS}})$ is positive semi-definite)

The OLS estimator $\widehat{\beta}_{\text{OLS}}$ is said to be the **efficient** estimator in the class of linear, unbiased estimators

This last result is known as the **Gauss-Markov theorem**, establishing that $\widehat{\beta}_{\text{OLS}}$ is the minimum variance linear unbiased estimator of the parameter vector β, or the 'best' linear unbiased estimator (BLUE), in the classical linear regression model with fixed regressors

Recall that $\widehat{\beta}_{\text{OLS}} = (X'X)^{-1}X'y = Ay$ where in this setting $A = (X'X)^{-1}X'$ is a non-stochastic $K \times n$ matrix

$\widehat{\beta}_{\text{OLS}} = Ay$ is thus a linear function of the random vector y

This is what is meant here by the term 'linear estimator'

$$\widehat{\beta}_{\text{OLS}} = (X'X)^{-1}X'y = Ay$$

Note that, as a linear function of the random vector y, $\widehat{\beta}_{\text{OLS}}$ is also a random vector

The random vector $\widehat{\beta}_{\text{OLS}}$ has an expected value equal to the true parameter vector ($\text{E}(\widehat{\beta}_{\text{OLS}}) = \beta$), where the expectation is taken over the (unspecified) distribution of the u errors, or equivalently (since $y = X\beta + u$ and $X\beta$ is non-stochastic) over the (unspecified) distribution of the y outcomes

Equivalently, since the true value of β is not stochastic, $\text{E}(\widehat{\beta}_{\text{OLS}} - \beta) = 0$, so the expected deviation from the true parameter vector is zero

The random vector $\widehat{\beta}_{\text{OLS}}$ has a variance ($\text{Var}(\widehat{\beta}_{\text{OLS}}) = \sigma^2(X'X)^{-1}$) which is smaller (in a matrix sense) than that of any other linear, unbiased estimator

Recalling that

$$\begin{aligned}\text{Var}(\widehat{\beta}_{\text{OLS}}) &= \text{E}[(\widehat{\beta}_{\text{OLS}} - \text{E}(\widehat{\beta}_{\text{OLS}}))(\widehat{\beta}_{\text{OLS}} - \text{E}(\widehat{\beta}_{\text{OLS}}))'] \\ &= \text{E}[(\widehat{\beta}_{\text{OLS}} - \beta)(\widehat{\beta}_{\text{OLS}} - \beta)'],\end{aligned}$$

the expected value of the squared deviation from the true parameter vector is minimised, within this class of estimators

The results that the OLS estimator $\widehat{\beta}_{\text{OLS}}$ is unbiased, and efficient in the class of linear, unbiased estimators, hold for any sample size $n > K$ (finite sample properties)

Although this does *not* mean that having more data would not be useful: *adding* more observations to the sample will generally *lower* $\text{Var}(\widehat{\beta}_{\text{OLS}})$, giving a more *precise* estimate of the parameter vector β, since the term $(X'X)^{-1}$ will generally be smaller (in a matrix sense)

What do these properties mean?

Imagine conducting a sequence of controlled experiments, with the same sample size n, the same values for the explanatory variables $x_1, ..., x_n$, and the same true parameter vector β

Different draws of the $u_1, ..., u_n$ errors (in the different experiments), or equivalently different draws of the $y_1, ..., y_n$ outcomes, would give different values for $\widehat{\beta}_{\text{OLS}}$, from the same (fixed) values of $x_1, ..., x_n$ and the same true parameter vector β

We would observe a distribution of the OLS estimates $\widehat{\beta}_{\text{OLS}}$ of the parameter vector β (we can do this using generated data in the context of Monte Carlo simulations)

If we repeated the experiment enough times, with independent draws of $u_1, ..., u_n$, (so that the sample mean converges to the expected value), then on average we would estimate the true value of the parameter vector, whatever the sample size n used in each experiment - this is the *unbiased* property

As well as being correct on average, the distribution of the OLS estimates would also have the lowest possible variance, in the class of linear, unbiased estimators, in these repeated experiments - this is the *efficiency* property

In practice, we typically observe only one sample, and we are typically not in a position to conduct repeated, controlled experiments

Nevertheless, if we want the estimator of the parameter vector that we compute from our sample to have no expected deviation from the true parameter vector, and to have the smallest possible variance within the class of linear estimators with this property, we know that the OLS estimator $\widehat{\beta}_{\text{OLS}}$ has these properties in the classical linear regression model (with fixed regressors)

For any $n \times 1$ random vector y, non-stochastic $K \times 1$ vector a, and non-stochastic $K \times n$ matrix B, the $K \times 1$ random vector $z = a + By$ has

$$\mathrm{E}(z) = a + B\mathrm{E}(y)$$

$$\mathrm{Var}(z) = B\mathrm{Var}(y)B'$$

Using these facts, we can easily show that

$$\mathrm{E}(\widehat{\beta}_{\mathrm{OLS}}) = A\mathrm{E}(y) = AX\beta = (X'X)^{-1}X'X\beta = \beta$$

$$\begin{aligned} \text{and } \mathrm{Var}(\widehat{\beta}_{\mathrm{OLS}}) &= A\mathrm{Var}(y)A' = A(\sigma^2 I)A' = \sigma^2 AA' \\ &= \sigma^2(X'X)^{-1}X'X(X'X)^{-1} = \sigma^2(X'X)^{-1} \end{aligned}$$

noting that $(X'X)^{-1}$ is a symmetric matrix, so that $[(X'X)^{-1}]' = (X'X)^{-1}$

We will not prove the Gauss-Markov theorem for this model, but rather show that similar results can be established in a version of the classical linear regression model that is more relevant for analyzing most kinds of data encountered in applied econometric research

From now on, some or all of the explanatory variables in the matrix X will be assumed to be stochastic, i.e. to be random variables (unless stated otherwise)

The classical linear regression model with stochastic regressors

The classical linear regression model with stochastic regressors can be stated as

$$y = X\beta + u$$

$$\mathrm{E}(u|X) = 0, \quad \mathrm{Var}(u|X) = \sigma^2 I$$

$$X \text{ has full rank (with probability one)}$$

or equivalently as

$$\mathrm{E}(y|X) = X\beta$$

$$\mathrm{Var}(y|X) = \sigma^2 I$$

$$X \text{ has full rank (with probability one)}$$

It may be instructive to derive this specification, starting with a linear model for observation i

$$\begin{aligned} y_i &= \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_K x_{Ki} + u_i \\ &= x_i'\beta + u_i \quad \text{for } i = 1, 2, ..., n \end{aligned}$$

First assume that the conditional expectation of y_i given x_i is linear, giving $\mathrm{E}(y_i|x_i) = x_i'\beta$, or equivalently $\mathrm{E}(u_i|x_i) = 0$. This is a **linear conditional expectation** assumption

Also assume that the conditional variance of y_i given x_i is common to all the observations $i = 1, 2, ..., n$, giving $\mathrm{Var}(y_i|x_i) = \sigma^2$, or equivalently $\mathrm{Var}(u_i|x_i) = \sigma^2$. This is a **conditional homoskedasticity** assumption

Further assume that the observations on (y_i, x_i) are **independent** over $i = 1, 2, ..., n$, so that

$$\mathrm{E}(y_i|x_1, x_2, ..., x_n) = \mathrm{E}(y_i|x_i) = x_i'\beta \text{ for all } i = 1, 2, ..., n$$

$$\text{or } \mathrm{E}(u_i|x_1, x_2, ..., x_n) = \mathrm{E}(u_i|x_i) = 0 \quad \text{for all } i = 1, 2, ..., n$$

and

$$\mathrm{Var}(y_i|x_1, x_2, ..., x_n) = \mathrm{Var}(y_i|x_i) = \sigma^2 \text{ for all } i = 1, 2, ..., n$$

$$\text{or } \mathrm{Var}(u_i|x_1, x_2, ..., x_n) = \mathrm{Var}(u_i|x_i) = \sigma^2 \text{ for all } i = 1, 2, ..., n$$

Independence also implies that

$$\text{Cov}(y_i, y_j | x_1, x_2, ..., x_n) = 0 \text{ for all } i \neq j$$

$$\text{or } \text{Cov}(u_i, u_j | x_1, x_2, ..., x_n) = 0 \text{ for all } i \neq j$$

Collecting these implications of independence gives the stronger forms of the linear conditional expectation assumption

$$\text{E}(y|X) = X\beta \quad \text{or} \quad \text{E}(u|X) = 0$$

and of the conditional homoskedasticity assumption

$$\text{Var}(y|X) = \sigma^2 I \quad \text{or} \quad \text{Var}(u|X) = \sigma^2 I$$

required for the classical linear regression model with stochastic regressors

Adding the assumption that X has full rank completes the specification

Since the $n \times K$ matrix X now contains random variables, some authors prefer to state the full rank assumption in the form $\text{rank}(X) = K$ with probability one

Although since this condition cannot be guaranteed with random draws from interesting distributions, even this is something of a fudge

Again we can note that the case of perfect multicollinearity (or rank deficiency) is easily avoided in practice

Stated in the form

$$\mathrm{E}(u_i|x_1, x_2, ..., x_n) = 0 \quad \text{for } i = 1, 2, ..., n$$

the linear conditional expectation assumption is sometimes referred to as 'strict exogeneity'

This implies that the error term u_i for observation i is uncorrelated with the explanatory variables in x_j for *all* observations $i, j = 1, 2, ..., n$, not *only* that u_i is uncorrelated with x_i

Similarly this can be shown to imply that the explanatory variables in x_i for observation i are uncorrelated with the error terms u_j for *all* observations $i, j = 1, 2, ..., n$, not *only* that x_i is uncorrelated with u_i

Notice that, in a time series setting, this strict exogeneity assumption rules out the presence of lagged dependent variables among the set of explanatory variables, since the dependent variable for observation $i-1$ is necessarily correlated with the error term for observation $i-1$

To show this more formally, we temporarily switch notation from $i = 1, ..., n$ to $t = 1, ..., T$ to denote the observations

Suppose that $K = 2$, with $x_{1t} = 1$ for all t (intercept term), and $x_{2t} = y_{t-1}$ (lagged dependent variable), giving the first-order autoregressive (AR(1)) model $y_t = \beta_1 + \beta_2 y_{t-1} + u_t$

$$y_t = \beta_1 + \beta_2 y_{t-1} + u_t = (1, y_{t-1}) \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} + u_t = x_t'\beta + u_t$$

$$\text{and } y_{t+1} = \beta_1 + \beta_2 y_t + u_{t+1} = (1, y_t) \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} + u_{t+1} = x_{t+1}'\beta + u_{t+1}$$

In this case, x_{t+1} *contains* y_t, so that

$$\begin{aligned} \mathrm{E}(y_t|x_1, ..., x_{t+1}, ..., x_T) &= y_t \\ &\neq \mathrm{E}(y_t|x_t) = x_t'\beta = \beta_1 + \beta_2 y_{t-1} \end{aligned}$$

We cannot maintain the independence assumption that was used to derive the classical linear regression model, and a different specification will be needed to handle dynamic models with lagged dependent variables in a time series setting

NB. We will not be able to obtain exact finite sample properties for the OLS estimator in such models, but we will still be able to obtain useful large sample properties

The assumption that

$$\text{Cov}(u_i, u_j | x_1, x_2, ..., x_n) = 0 \text{ for all } i \neq j$$

also rules out the case of serially correlated errors in time series models, where for example we may want to specify that

$$\text{Cov}(u_t, u_{t-1} | x_1, x_2, ..., x_T) \neq 0$$

allowing for forms of temporal dependence or persistence in the error terms

In a cross-section setting, the strict exogeneity assumption rules out any correlation between the error terms for different observations, or any 'cross-section dependence'

Combined with the assumption of conditional homoskedasticity, this is sometimes stated more strongly as the requirement that, conditional on the explanatory variables in X, the error terms for different observations are independently and identically distributed (iid errors)

Although notice that we have not specified the form of the distribution function from which these error terms are drawn

And, in particular, we have *not* assumed that the errors are drawn from a normal distribution

We now consider the **finite sample properties** of the OLS estimator in the setting of the classical linear regression model with stochastic regressors

We first use the facts that, for any $n \times 1$ vector y that is stochastic (i.e. random) conditional on X, any $K \times 1$ vector a that is non-stochastic (i.e. known) conditional on X, and any $K \times n$ matrix B that is non-stochastic conditional on X, the $K \times 1$ vector $z = a + By$ is stochastic conditional on X, with

$$\mathrm{E}(z|X) = a + B\mathrm{E}(y|X)$$

$$\mathrm{Var}(z|X) = B\mathrm{Var}(y|X)B'$$

Recall that $\widehat{\beta}_{\text{OLS}} = (X'X)^{-1}X'y = Ay$

Since $A = (X'X)^{-1}X'$ is known (i.e. not stochastic) conditional on X, while y is a random vector conditional on X, we have that $\widehat{\beta}_{\text{OLS}}$ is a random vector conditional on X

The conditional expectation of $\widehat{\beta}_{\text{OLS}}$ is

$$\text{E}(\widehat{\beta}_{\text{OLS}}|X) = A\text{E}(y|X) = AX\beta = (X'X)^{-1}X'X\beta = \beta$$

since $A = (X'X)^{-1}X'$ is not stochastic conditional on X

This shows that the OLS estimator $\widehat{\beta}_{\text{OLS}}$ is **unbiased**, conditional on the realized values of the stochastic regressors X observed in our sample

Again this implies that $\mathrm{E}(\widehat{\beta}_{\mathrm{OLS}} - \beta | X) = 0$, so that given the realized values of the regressors, we have zero expected deviation from the true parameter vector

The expectation is taken over the (unspecified) conditional distribution of $u|X$, or equivalently (since $y = X\beta + u$ and $X\beta$ is not stochastic conditional on X) over the (unspecified) conditional distribution of $y|X$

The thought experiment here is that we fix the value of X and calculate $\widehat{\beta}_{\mathrm{OLS}}$ for many different (independent) draws of $u|X$ or $y|X$

On average we estimate the true value of the parameter vector

We can also show that $\widehat{\beta}_{\text{OLS}}$ is unbiased in a weaker, unconditional sense, using the **Law of Iterated Expectations**

$$\text{E}(z) = \text{E}[\text{E}(z|X)]$$

where the expectation E[.] is taken over the distribution of X

This gives the unconditional expectation of $\widehat{\beta}_{\text{OLS}}$ as

$$\text{E}(\widehat{\beta}_{\text{OLS}}) = \text{E}[\text{E}(\widehat{\beta}_{\text{OLS}}|X)] = \text{E}[\beta] = \beta$$

since the true parameter vector β is not stochastic

This unconditional unbiasedness property is more relevant in most economic contexts, since it is rarely meaningful to think in terms of fixing the values of the explanatory variables (this is why we want to treat the regressors as stochastic)

In practice, different samples we could use to estimate β will contain different values of the explanatory variables X, as well as different values of the outcome variable y

Again we can state this property as $\mathrm{E}(\widehat{\beta}_{\mathrm{OLS}} - \beta) = 0$, so the expected deviation from the true parameter vector is zero

The conditional variance of $\widehat{\beta}_{\text{OLS}}$ is

$$\begin{aligned}\text{Var}(\widehat{\beta}_{\text{OLS}}|X) &= A\text{Var}(y|X)A' = A(\sigma^2 I)A' = \sigma^2 AA' \\ &= \sigma^2(X'X)^{-1}X'X(X'X)^{-1} = \sigma^2(X'X)^{-1}\end{aligned}$$

noting that $(X'X)^{-1}$ is a symmetric matrix, as before

The Law of Iterated Expectations can be used to show that the unconditional variance is $\text{Var}(\widehat{\beta}_{\text{OLS}}) = \sigma^2\text{E}[(X'X)^{-1}]$

But the expression for the conditional expectation $\text{Var}(\widehat{\beta}_{\text{OLS}}|X) = \sigma^2(X'X)^{-1}$ is more useful here

This will be used in the development of hypothesis tests about the true value of (elements of) the parameter vector, which are conducted conditional on the observed values of the regressors in our sample

We also have a conditional version of the **Gauss-Markov theorem**, stating that

$$\text{Var}(\widehat{\beta}_{\text{OLS}}|X) \leqslant \text{Var}(\widetilde{\beta}|X)$$

for any other unbiased estimator $\widetilde{\beta}$ that is a linear function of the random vector y conditional on X

Given the realized values of X, the OLS estimator $\widehat{\beta}_{\text{OLS}}$ is **efficient** in the class of linear, unbiased estimators

To prove this version of the Gauss-Markov theorem, we let $\widetilde{\beta} = \widetilde{A}y$ for a $K \times n$ matrix $\widetilde{A}$ that is non-stochastic conditional on X (i.e. $\widetilde{\beta}$ is a linear function of y conditional on X)

For $\widetilde{\beta}$ to be unbiased conditional on X, we require $\mathrm{E}(\widetilde{\beta}|X) = \widetilde{A}\mathrm{E}(y|X) = \widetilde{A}X\beta = \beta$

Hence we require $\widetilde{A}X = I$ (here a $K \times K$ identity matrix)

We also have $\mathrm{Var}(\widetilde{\beta}|X) = \widetilde{A}\mathrm{Var}(y|X)\widetilde{A}' = \sigma^2\widetilde{A}\widetilde{A}'$

Write $\widetilde{A} = (A+D)$ where $A = (X'X)^{-1}X'$ and the $K \times n$ matrix $D = \widetilde{A} - A$

Consider $\widetilde{A}X = (A + D)X = AX + DX$

We know that $AX = (X'X)^{-1}X'X = I$, which implies $\widetilde{A}X = I + DX$

For $\widetilde{\beta}$ to be unbiased conditional on X, we require $\widetilde{A}X = I$, which implies $DX = 0$ (a $K \times K$ matrix with every element zero)

Note that $DX = 0$ implies $DX(X'X)^{-1} = DA' = 0$ (using the symmetry of $(X'X)^{-1}$), which in turn implies that $[DA']' = AD' = 0$

Now consider $\widetilde{A}\widetilde{A}' = (A+D)(A+D)' = AA' + AD' + DA' + DD'$

If $\widetilde{\beta}$ is unbiased conditional on X, we have $AD' = DA' = 0$, and this simplifies to $\widetilde{A}\widetilde{A}' = AA' + DD'$

Now since $\mathrm{Var}(\widetilde{\beta}|X) = \sigma^2\widetilde{A}\widetilde{A}'$, we have $\mathrm{Var}(\widetilde{\beta}|X) = \sigma^2 AA' + \sigma^2 DD'$

Recall that $\mathrm{Var}(\widehat{\beta}_{\mathrm{OLS}}|X) = \sigma^2(X'X)^{-1} = \sigma^2 AA'$

So we have $\mathrm{Var}(\widetilde{\beta}|X) = \mathrm{Var}(\widehat{\beta}_{\mathrm{OLS}}|X) + \sigma^2 DD'$

For any $K \times n$ matrix D, the matrix DD' is positive semi-definite

Since $\sigma^2 = \mathrm{Var}(u_i|X) > 0$, we also have $\sigma^2 DD'$ is positive semi-definite

Thus $\mathrm{Var}(\widetilde{\beta}|X) - \mathrm{Var}(\widehat{\beta}_{\mathrm{OLS}}|X)$ is positive semi-definite, which is what is meant by the matrix inequality statement $\mathrm{Var}(\widetilde{\beta}|X) \geqslant \mathrm{Var}(\widehat{\beta}_{\mathrm{OLS}}|X)$

Two implications of this result can also be noted

$$\mathrm{Var}(\widehat{\beta}_{\mathrm{OLS}}|X) = \begin{pmatrix} \mathrm{Var}(\widehat{\beta}_1|X) & \mathrm{Cov}(\widehat{\beta}_1, \widehat{\beta}_2|X) & \cdots & \mathrm{Cov}(\widehat{\beta}_1, \widehat{\beta}_K|X) \\ \mathrm{Cov}(\widehat{\beta}_2, \widehat{\beta}_1|X) & \mathrm{Var}(\widehat{\beta}_2|X) & \cdots & \mathrm{Cov}(\widehat{\beta}_2, \widehat{\beta}_K|X) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}(\widehat{\beta}_K, \widehat{\beta}_1|X) & \mathrm{Cov}(\widehat{\beta}_K, \widehat{\beta}_2|X) & \cdots & \mathrm{Var}(\widehat{\beta}_K|X) \end{pmatrix}$$

where $\widehat{\beta}_k$ denotes the k^{th} element of $\widehat{\beta}_{\mathrm{OLS}}$ for $k = 1, ..., K$

Variance matrices are symmetric, since $\mathrm{Cov}(\widehat{\beta}_j, \widehat{\beta}_k|X) = \mathrm{Cov}(\widehat{\beta}_k, \widehat{\beta}_j|X)$

Similarly $\text{Var}(\widetilde{\beta}|X)$ and $\text{Var}(\widetilde{\beta}|X) - \text{Var}(\widehat{\beta}_{\text{OLS}}|X)$ are symmetric matrices

A symmetric matrix that is positive semi-definite has non-negative numbers on its main diagonal

So $\text{Var}(\widetilde{\beta}|X) - \text{Var}(\widehat{\beta}_{\text{OLS}}|X)$ positive semi-definite implies that the diagonal elements $\text{Var}(\widetilde{\beta}_k|X) - \text{Var}(\widehat{\beta}_k|X) \geqslant 0$ or $\text{Var}(\widetilde{\beta}_k|X) \geqslant \text{Var}(\widehat{\beta}_k|X)$ for each $k = 1, ..., K$

For *each element* of the parameter vector, the OLS estimator has the smallest variance in the class of estimators that are linear and unbiased, conditional on X

Any linear combination of the β_k parameters can be expressed in the form $\theta = h'\beta$, for some non-stochastic $K \times 1$ vector h

For example, $\beta_2 - \beta_1$ is obtained using $h' = (-1, 1, 0, ..., 0)$ so that $h'\beta = -\beta_1 + \beta_2 = \beta_2 - \beta_1$

The estimator $\widehat{\theta}_{\text{OLS}}$ based on $\widehat{\beta}_{\text{OLS}}$ is $\widehat{\theta}_{\text{OLS}} = h'\widehat{\beta}_{\text{OLS}}$

The estimator $\widetilde{\theta}$ based on $\widetilde{\beta}$ is $\widetilde{\theta} = h'\widetilde{\beta}$

We have $\text{Var}(\widehat{\theta}_{\text{OLS}}|X) = h'\text{Var}(\widehat{\beta}_{\text{OLS}}|X)h$ and $\text{Var}(\widetilde{\theta}|X) = h'\text{Var}(\widetilde{\beta}|X)h$

So that $\text{Var}(\widetilde{\theta}|X) - \text{Var}(\widehat{\theta}_{\text{OLS}}|X) = h'[\text{Var}(\widetilde{\beta}|X) - \text{Var}(\widehat{\beta}_{\text{OLS}}|X)]h$

For any $K \times 1$ vector h and positive semi-definite matrix G, the scalar $h'Gh \geqslant 0$

Hence $\text{Var}(\widetilde{\theta}|X) - \text{Var}(\widehat{\theta}_{\text{OLS}}|X) \geqslant 0$ or $\text{Var}(\widetilde{\theta}|X) \geqslant \text{Var}(\widehat{\theta}_{\text{OLS}}|X)$

For *any linear combination* of the β_k parameters, the OLS estimator has the smallest variance in the class of estimators that are linear and unbiased, conditional on X

Our previous result for individual elements of the parameter vector β can also be obtained as a special case of this result; for example, using $h' = (1, 0, ..., 0)$ gives $h'\beta = \beta_1$

Fitted values

The fitted value for observation i in our sample is $\widehat{y}_i = x_i'\widehat{\beta}_{\text{OLS}}$

The vector of fitted values for all n observations in our sample is

$$\widehat{y} = X\widehat{\beta}_{\text{OLS}} = X(X'X)^{-1}X'y = Py$$

where the $n \times n$ matrix $P = X(X'X)^{-1}X'$

The matrix P has the property that $PX = X$

The matrix P is sometimes called the 'projection matrix'

Residuals

The residual for observation i is $\widehat{u}_i = y_i - x_i'\widehat{\beta}_{\text{OLS}} = y_i - \widehat{y}_i$

The vector of residuals for all n observations in our sample is

$$\begin{aligned}\widehat{u} &= y - X\widehat{\beta}_{\text{OLS}} = y - \widehat{y} \\ &= y - X(X'X)^{-1}X'y = (I - P)y = My\end{aligned}$$

where the $n \times n$ matrix $M = I - P = I - X(X'X)^{-1}X'$

The matrix M has the property that $MX = 0$

The matrix M is sometimes called the 'annihilator'

The K normal equations used to obtain the OLS estimator imply that, for the OLS residuals in our sample, we have $X'\widehat{u} = 0$, or

$$\begin{pmatrix} \sum_{i=1}^{n} x_{1i}\widehat{u}_i \\ \sum_{i=1}^{n} x_{2i}\widehat{u}_i \\ \vdots \\ \sum_{i=1}^{n} x_{Ki}\widehat{u}_i \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

This implies that, by construction, the OLS residuals $\widehat{u}_1, ..., \widehat{u}_n$ in our sample are uncorrelated with each of the K explanatory variables included in the model

This sample property of the OLS residuals follows directly from the definition of the OLS estimator, and tells us nothing about the validity of the linear conditional expectation assumption

$$\mathrm{E}(u_i|x_i) = \mathrm{E}(u_i|x_{1i}, x_{2i}, ..., x_{Ki}) = 0$$

or the (stronger) strict exogeneity assumption

$$\mathrm{E}(u_i|x_1, x_2, ..., x_n) = 0$$

made in the classical linear regression model with stochastic regressors

Estimation of σ^2

The variance parameter σ^2 is usually unknown

For the result that $\text{Var}(\widehat{\beta}_{\text{OLS}}|X) = \sigma^2(X'X)^{-1}$ to be useful in practice, we need an estimator for σ^2

The OLS estimator is

$$\widehat{\sigma}^2_{\text{OLS}} = \frac{\widehat{u}'\widehat{u}}{n-K} = \frac{\sum_{i=1}^{n} \widehat{u}_i^2}{n-K}$$

where $\widehat{u} = y - X\widehat{\beta}_{\text{OLS}}$ is the vector of OLS residuals

The estimator $\widehat{\sigma}^2_{\text{OLS}}$ can be shown to be an unbiased estimator of σ^2 in the classical linear regression models

With stochastic regressors, this holds both conditional on X, and also unconditionally

Recall that in the special case with $K = 1$ and $x_{1i} = 1$ for $i = 1, 2, ..., n$ (that is, in the model $y_i = \beta_1 + u_i$), the OLS estimator $\widehat{\beta}_1 = \overline{y}_i$, so that the OLS residuals $\widehat{u}_i = y_i - \overline{y}_i$

In this case, $\widehat{\sigma}^2_{\text{OLS}}$ simplifies to the sample variance $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y}_i)^2$

For independent observations with $\text{E}(y_i) = \mu$ and $\text{Var}(y_i) = \sigma^2$, it follows that the sample variance is an unbiased estimator of σ^2

OLS and Maximum Likelihood

The OLS estimator that we derived by minimizing the sum of the squared errors can also be obtained as a (conditional) Maximum Likelihood (ML) estimator in a particular case of the linear model

If the $n \times 1$ random vector y has the normal distribution $y \sim \mathrm{N}(\mu, \Sigma)$, its probability density function is

$$f_y(y = a) = f_y(a) = (2\pi)^{\frac{-n}{2}} |\Sigma|^{\frac{-1}{2}} \exp\left(\frac{-w}{2}\right)$$

where $w = u'\Sigma^{-1}u$, $u = a - \mu$ and $|\Sigma|^{\frac{-1}{2}} = \frac{1}{\sqrt{\det(\Sigma)}}$

If we add to our specification of the classical linear regression model with stochastic regressors the *further* assumption that $u|X \sim \mathrm{N}(0, \sigma^2 I)$ or, equivalently, that $y|X \sim \mathrm{N}(X\beta, \sigma^2 I)$, we obtain the conditional probability density function

$$f_{y|X}(a|X) = (2\pi)^{\frac{-n}{2}}(\sigma^2)^{\frac{-n}{2}} \exp\left(\frac{-u'u}{2\sigma^2}\right)$$

where $u = a - X\beta$, $\Sigma^{-1} = \left(\frac{1}{\sigma^2}\right) I$, $w = \left(\frac{u'u}{\sigma^2}\right)$ and $|\Sigma| = (\sigma^2)^n$

Viewed as a conditional density function, the argument is a, with both the realized value of X and the true values of the parameters β and σ^2 taken as given

But we can also view the expression $(2\pi)^{\frac{-n}{2}}(\sigma^2)^{\frac{-n}{2}}\exp\left(\frac{-u'u}{2\sigma^2}\right)$ as a function of the parameters β and σ^2, with $u = y - X\beta$, and the data on y and X now taken as given

This gives the (conditional) **likelihood function** for the sample data on (y, X) as

$$\mathcal{L}(\beta, \sigma^2) = (2\pi)^{\frac{-n}{2}}(\sigma^2)^{\frac{-n}{2}}\exp\left(\frac{-u'u}{2\sigma^2}\right)$$

(Conditional) maximum likelihood estimators find the values of the parameters β and σ^2 that maximize this (conditional) likelihood function given the sample data

Or, equivalently, which maximize the (conditional) log-likelihood function

$$L(\beta, \sigma^2) = \ln \mathcal{L}(\beta, \sigma^2) = -\left(\frac{n}{2}\right) \ln 2\pi - \left(\frac{n}{2}\right) \ln(\sigma^2) - \left(\frac{1}{2\sigma^2}\right) u'u$$

with $u = y - X\beta$

Notice that only the final term in the log-likelihood function depends on β

Maximizing $L(\beta, \sigma^2)$ with respect to β is thus equivalent to minimizing the sum of squared errors $u'u = \sum_{i=1}^{n} u_i^2$ with respect to β

Hence the (conditional) maximum likelihood estimator $\widehat{\beta}_{\mathrm{ML}}$ in the classical linear regression model with normally distributed errors is identical to the OLS estimator $\widehat{\beta}_{\mathrm{OLS}}$

Strictly this is a conditional maximum likelihood estimator, since we have only specified the conditional distribution of $y|X$, or equivalently the conditional distribution of $u|X$

But this is commonly shortened to maximum likelihood estimator

[Starting from the joint density of (y, X), $f_{y,X}$, we can always factorise this joint density into the conditional density $f_{y|X}$ times the marginal density f_X

The distinction between (true) maximum likelihood estimators based on $f_{y,X}$ and conditional maximum likelihood estimators based on $f_{y|X}$ only matters in (unusual) cases where $f_{y|X}$ and f_X have parameters in common]

OLS and Generalized Method of Moments

The OLS estimator can also be derived as a (Generalized) Method of Moments (GMM) estimator in the linear model formulated as

$$y_i = x_i'\beta + u_i \qquad \text{with} \quad \mathrm{E}(u_i) = 0 \quad \text{and} \quad \mathrm{E}(x_i u_i) = 0 \quad \text{for } i = 1, ..., n$$

Method of Moments estimators find the value of β that sets the **sample analogue** of the population moment condition $\mathrm{E}(x_i u_i) = 0$ equal to zero

The sample analogue of an expected value is the sample mean

So we find the value of β that sets $\frac{1}{n}\sum_{i=1}^{n} x_i u_i(\beta) = 0$, where $u_i(\beta) = y_i - x_i'\beta$

But $\frac{1}{n}\sum_{i=1}^{n} x_i u_i(\beta) = \left(\frac{1}{n}\right) X'u(\beta)$

And the value of β that sets $X'u(\beta) = 0$ is again the OLS estimator $\widehat{\beta}_{\text{OLS}}$

So we also have $\widehat{\beta}_{\text{OLS}} = \widehat{\beta}_{\text{GMM}}$ in this linear model

We will look at properties of ML and GMM estimators in more general settings later in the course

Both approaches produce an estimator with some desirable properties in versions of the linear model