

Lecture 2: Estimating the Classical Linear Regression Model

by Ordinary Least Squares (OLS)

(Hayashi, pp. 15-21)

1. Recap: The Data and the Model to be Estimated
2. Computing the Least Squares Coefficients (**b**)
3. Some related concepts (OLS residuals, s^2 , R^2 , sampling error)

1. Recap: The Data and the Model to be Estimated

a) The **Data** consist of an n -element column vector of outcomes, \mathbf{y} (for example the log earnings of 50,000 persons in a cross-sectional survey), and an $n \times K$ matrix of independent variables, \mathbf{X} , where each row of \mathbf{X} gives the independent variables for observation (person) $i=1, \dots, n$, and each column of \mathbf{X} gives the information for variable $j = 1, \dots, K$:

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}; \quad \mathbf{X} = \begin{bmatrix} \mathbf{x}'_1 \\ \mathbf{x}'_2 \\ \vdots \\ \mathbf{x}'_n \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1K} \\ x_{21} & x_{22} & \cdots & x_{2K} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nK} \end{bmatrix}$$

b) The *Model*. We *suppose* that these data were generated by the following underlying process:

i) Some process, possibly random, determines \mathbf{X} . The only restrictions we assume on this process are that \mathbf{X} contains a column of ones, and that the rank of the $n \times K$ data matrix, \mathbf{X} , is K with probability one [*no multicollinearity*].

ii) For each observation, a random error, ε_i , is drawn independently from the *same* underlying univariate probability distribution, which has mean zero. [this gives us *strict exogeneity* and *homoskedasticity*, i.e. $E(\boldsymbol{\varepsilon} | \mathbf{X}) = \mathbf{0}$ and $E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}' | \mathbf{X}) = \sigma^2 \mathbf{I}_n$, where $\boldsymbol{\varepsilon}$ is the $n \times 1$ vector of errors].

iii) For each observation, y_i is determined by: $y_i = \sum_{k=1}^K \beta_k x_{ik} + \varepsilon_i$. Thus, for the data set as a whole:

$$\underbrace{\mathbf{y}}_{(n \times 1)} = \underbrace{\mathbf{X}}_{(n \times K)} \underbrace{\boldsymbol{\beta}}_{(K \times 1)} + \underbrace{\boldsymbol{\varepsilon}}_{(n \times 1)}. \quad [\text{This is } \textit{linearity}].$$

The above model is our *maintained hypothesis*—for the next 3 lectures we simply assume it is true.

This model, however, contains $K + 1$ unknown parameters: the vector β and the variance of the distribution of ε : σ^2 .

How, then, might we estimate the values of these true, underlying parameters, β and σ^2 , from the data available to us, y and \mathbf{X} ?

Well, one procedure would be what we call OLS, namely to:

1. Choose b_1, \dots, b_K to minimize the sum of squared residuals:

$$\text{Min} \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \sum_{k=1}^K b_k x_{ik})^2,$$

2. Calculate s^2 as: $s^2 = \sum_{i=1}^n e_i^2 / (n - K)$.

Now we will actually do this (Solve minimization problem in 1 above).

Note—the *mechanics* of estimation only require *no multicollinearity*).

2. Computing the Least Squares Estimates

First, some important distinctions/notation:

$\boldsymbol{\beta}$ ($K \times 1$): The true (unknown) coefficient vector.

$\tilde{\boldsymbol{\beta}}$ ($K \times 1$): A candidate estimated coefficient vector.

\mathbf{b} ($K \times 1$): The OLS coefficient vector (estimated by minimizing the SSR).

To compute the least-squares estimates, \mathbf{b} , note first that the sum of squared residuals for *an arbitrary coefficient vector*, $\tilde{\boldsymbol{\beta}}$, can be written:

$$\text{SSR}(\tilde{\boldsymbol{\beta}}) = \tilde{\mathbf{e}}' \tilde{\mathbf{e}} = (\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}})$$

Rewrite this to make it easier to differentiate wrt $\tilde{\boldsymbol{\beta}}$:

$$\begin{aligned}
 \text{SSR}(\tilde{\boldsymbol{\beta}}) &= \underbrace{(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}})'}_{(1 \times n)} \underbrace{(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}})}_{(n \times 1)} \\
 &= \underbrace{(\mathbf{y}' - \tilde{\boldsymbol{\beta}}'\mathbf{X}')}_{1 \times n} \underbrace{(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}})}_{n \times 1} \quad (\text{since } (\mathbf{X}\tilde{\boldsymbol{\beta}})' = \tilde{\boldsymbol{\beta}}'\mathbf{X}') \\
 &= \mathbf{y}'\mathbf{y} - \tilde{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} - \mathbf{y}'\mathbf{X}\tilde{\boldsymbol{\beta}} + \tilde{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\tilde{\boldsymbol{\beta}} \\
 &= \mathbf{y}'\mathbf{y} - 2\mathbf{y}'\mathbf{X}\tilde{\boldsymbol{\beta}} + \tilde{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\tilde{\boldsymbol{\beta}} \quad (\text{since } \tilde{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} \text{ and } \mathbf{y}'\mathbf{X}\tilde{\boldsymbol{\beta}} \text{ are both scalars}) \quad (1)
 \end{aligned}$$

The first term in (1) does not depend on $\tilde{\boldsymbol{\beta}}$.

The second is just a cross-product of two K -element vectors, $\mathbf{a}'\tilde{\boldsymbol{\beta}}$, where

$\mathbf{a}' = \underbrace{-2\mathbf{y}'}_{(1 \times n)} \underbrace{\mathbf{X}}_{(n \times K)}$, so its derivative wrt $\tilde{\boldsymbol{\beta}}$ is just:

$$\mathbf{a} = -2\mathbf{X}'\mathbf{y} \quad (K \times 1)$$

The third term is a quadratic form, $\tilde{\boldsymbol{\beta}}' \mathbf{A} \tilde{\boldsymbol{\beta}}$, where $\mathbf{A} = \mathbf{X}'\mathbf{X}$ and is therefore symmetric. So its derivative wrt $\tilde{\boldsymbol{\beta}}$ is just:

$$2\mathbf{A}\tilde{\boldsymbol{\beta}}, \text{ or } 2\mathbf{X}'\mathbf{X}\tilde{\boldsymbol{\beta}} . (K \times 1)$$

So (using \mathbf{b} to denote the OLS estimator) the FOC for a minimum of the SSR are:

$$\underbrace{-2\mathbf{X}'\mathbf{y}}_{(K \times n)(n \times 1)} + \underbrace{2\mathbf{X}'\mathbf{X}\mathbf{b}}_{(K \times K)(K \times 1)} = \mathbf{0}, \text{ or}$$

$$\underbrace{\mathbf{X}'\mathbf{X}}_{K \times K} \underbrace{\mathbf{b}}_{K \times 1} = \underbrace{\mathbf{X}'}_{K \times n} \underbrace{\mathbf{y}}_{n \times 1} \quad (2)$$

(2) is a system of K linear equations, for K unknowns, i.e. the elements of \mathbf{b} . These are sometimes called the *normal equations*.

Provided \mathbf{X} is of full column rank (Assumption 3—no multicollinearity), $\mathbf{X}'\mathbf{X}$ will be positive definite, and therefore nonsingular. This allows us to invert $\mathbf{X}'\mathbf{X}$, to get a closed-form, unique solution for the OLS coefficient vector:

The OLS estimator:

$$\underbrace{\mathbf{b}}_{K \times 1} = \underbrace{(\mathbf{X}'\mathbf{X})^{-1}}_{K \times K} \underbrace{\mathbf{X}'}_{K \times n} \underbrace{\mathbf{y}}_{n \times 1}$$

Note: Sometimes it will be handy to write this as: $\underbrace{\mathbf{b}}_{K \times 1} = \underbrace{\mathbf{A}}_{K \times n} \underbrace{\mathbf{y}}_{n \times 1}$,

where $\underbrace{\mathbf{A}}_{K \times n} = \underbrace{(\mathbf{X}'\mathbf{X})^{-1}}_{K \times K} \underbrace{\mathbf{X}'}_{K \times n}$

3. Some related concepts

The *fitted* (“predicted”) *value* for observation i : $\hat{y}_i = \mathbf{x}'_i \mathbf{b}$

The *OLS residual* for observation i : $e_i = y_i - \hat{y}_i = y_i - \mathbf{x}'_i \mathbf{b}$

Note this is NOT the same as the error term, ε_i .

The $(n \times 1)$ vector of fitted (“predicted”) values: \mathbf{Xb}

The $(n \times 1)$ vector of OLS residuals: $\mathbf{e} = \mathbf{y} - \mathbf{Xb}$

The *sum of squared of OLS residuals*: $SSR = \mathbf{e}'\mathbf{e}$

(Note: by construction, the vector of OLS residuals satisfies

$$\underbrace{\mathbf{X}'}_{K \times n} \underbrace{\mathbf{e}}_{n \times 1} = \mathbf{0}; \quad \text{i.e. the vector } \mathbf{e} \text{ is orthogonal to the vector } \mathbf{x}_k \text{ for every } k, \text{ i.e.}$$

for every \mathbf{x} variable. Why? Just re-arrange the normal equations, (2).)

$$\text{The } \mathbf{OLS} \text{ estimate of } \sigma^2 \equiv s^2 = \frac{SSR}{n-K} = \frac{\mathbf{e}'\mathbf{e}}{n-K}$$

Why does it differ from its 'sample analogue'? By construction, \mathbf{e} has to satisfy the K constraints given by (2); this limits its possible variability.

$$\begin{aligned}
\text{The } \textit{sampling error}, (\mathbf{b} - \boldsymbol{\beta}), &= \underbrace{(\mathbf{X}'\mathbf{X})^{-1}}_{K \times K} \underbrace{\mathbf{X}'}_{K \times n} \underbrace{y}_{n \times 1} - \boldsymbol{\beta} \\
&= \underbrace{(\mathbf{X}'\mathbf{X})^{-1}}_{K \times K} \underbrace{\mathbf{X}'}_{K \times n} \underbrace{(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon})}_{n \times 1} - \boldsymbol{\beta} \\
&= \underbrace{(\mathbf{X}'\mathbf{X})^{-1}}_{K \times K} \underbrace{(\mathbf{X}'\mathbf{X})}_{K \times K} \underbrace{\boldsymbol{\beta}}_{K \times 1} + \underbrace{(\mathbf{X}'\mathbf{X})^{-1}}_{K \times K} \underbrace{\mathbf{X}'\boldsymbol{\varepsilon}}_{K \times 1} - \boldsymbol{\beta} \\
&= \underbrace{\boldsymbol{\beta}}_{K \times 1} + \underbrace{(\mathbf{X}'\mathbf{X})^{-1}}_{K \times K} \underbrace{\mathbf{X}'\boldsymbol{\varepsilon}}_{K \times 1} - \underbrace{\boldsymbol{\beta}}_{K \times 1} \\
&= \underbrace{(\mathbf{X}'\mathbf{X})^{-1}}_{K \times K} \underbrace{\mathbf{X}'\boldsymbol{\varepsilon}}_{K \times 1} = \underbrace{\mathbf{A}}_{(K \times n)(n \times 1)} \underbrace{\boldsymbol{\varepsilon}}_{n \times 1}
\end{aligned}$$

Decomposing the sum of squares ($\mathbf{y}'\mathbf{y}$):

$$\begin{aligned}\mathbf{y}'\mathbf{y} &= (\hat{\mathbf{y}} + \mathbf{e})'(\hat{\mathbf{y}} + \mathbf{e}) = \hat{\mathbf{y}}'\hat{\mathbf{y}} + 2\hat{\mathbf{y}}'\mathbf{e} + \mathbf{e}'\mathbf{e} \\ &= \hat{\mathbf{y}}'\hat{\mathbf{y}} + 2\mathbf{b}'\mathbf{X}'\mathbf{e} + \mathbf{e}'\mathbf{e}.\end{aligned}$$

But, as already noted, $\mathbf{X}'\mathbf{e} = \mathbf{0}$ by construction. Thus,

$$\mathbf{y}'\mathbf{y} = \hat{\mathbf{y}}'\hat{\mathbf{y}} + \mathbf{e}'\mathbf{e} \quad (3)$$

The sum of squared y 's just equals the sum of squares of the predicted y 's plus the sum of squares of the OLS residuals. We can ignore the covariance between between \hat{y}_i and e_i , because it's zero *by construction* (not by assumption).

This decomposition allows us to define a simple goodness-of-fit measure for our regression, as follows:

The **Centered R^2** , or *coefficient of determination*, is the share of the total variance of \mathbf{y} that is accounted for by all the variables in \mathbf{X} . To justify this measure, we express (3) in terms of variances:

By the formula for variance:

$$Var(\mathbf{y}) = \frac{\mathbf{y}'\mathbf{y} - \bar{y}^2}{n}, \text{ where } \bar{y} \text{ is the sample mean of } y, \text{ and}$$

$$Var(\hat{\mathbf{y}}) = \frac{\hat{\mathbf{y}}'\hat{\mathbf{y}} - \bar{y}^2}{n} \text{ (which uses the fact that the } \hat{y} \text{ and } y \text{ have the same mean).}$$

Rearranging, it follows that $\mathbf{y}'\mathbf{y} = nVar(\mathbf{y}) + \bar{y}^2$ and $\hat{\mathbf{y}}'\hat{\mathbf{y}} = nVar(\hat{\mathbf{y}}) + \bar{y}^2$.

Substituting these into (3), simplifying (and using the fact that $E(e)=0$) yields:

$Var(\mathbf{y}) = Var(\hat{\mathbf{y}}) + Var(\mathbf{e})$. It now makes sense to define:

$$R^2 = \frac{Var(\hat{\mathbf{y}})}{Var(\mathbf{y})} = 1 - \frac{Var(\mathbf{e})}{Var(\mathbf{y})}.$$