

Deriving OLS Estimates

Want a line that best summarises the relationship suggested by the data.

a) 2 Variable Model

If we wish to fit a (straight) line through N observations, then think of this as giving a predicted value for the dependent variable conditional on the observed value of the independent variable

$$\hat{y}_i = \hat{b}_0 + \hat{b}_1 X_i$$

The difference between the predicted and actual value is called the **residual**

$$u = y_i - \hat{y}_i$$

The OLS principle says choose \hat{b}_0 and \hat{b}_1 to minimise the sum of squared residuals (avoids problems of negative residuals being offset by positive residuals, larger residuals receive more weight when squared and so any line with large residuals will be far from zero in the sum)

$$S = u_1^2 + u_2^2 + \dots + u_N^2 = \sum_{i=1}^N u_i^2 = \sum_{i=1}^N (y_i - \hat{y}_i)^2$$

sub. in $\hat{y}_i = \hat{b}_0 + \hat{b}_1 X_i$

$$S = \sum_{i=1}^N (Y_i - \hat{b}_0 - \hat{b}_1 X_i)^2$$

Find values of \hat{b}_0 and \hat{b}_1 which minimise this sum, using simple calculus rules

$$1) \quad \frac{dS}{d\hat{b}_0} = 0 \quad \text{and} \quad 2) \quad \frac{dS}{d\hat{b}_1} = 0$$

$$(1) \quad \frac{\partial S}{\partial \hat{b}_0} = 0 \Rightarrow 2N\hat{b}_0 - 2\sum Y_i + 2\hat{b}_1 \sum X_i = 0$$

and

$$(2) \quad \frac{\partial S}{\partial \hat{b}_1} = 0 \Rightarrow 2\hat{b}_1 \sum X_i^2 - 2\sum X_i Y_i + 2\hat{b}_0 \sum X_i = 0$$

(1) and (2) are known as the **normal equations**

Using the fact that the sample means of Y and X

$$\bar{Y}_i = \frac{\sum_{i=1}^N y_i}{N} \Leftrightarrow N \bar{Y}_i = \sum_{i=1}^N y_i \quad \text{and}$$

$$\bar{X}_i = \frac{\sum_{i=1}^N x_i}{N} \Leftrightarrow N \bar{X}_i = \sum_{i=1}^N x_i$$

can re-write (1) as

$$2N \hat{b}_0 - 2N \bar{Y} + 2\hat{b}_1 N \bar{X} = 0$$

and so obtain the formula to calculate the OLS estimate of the intercept

$$\hat{b}_0 = \bar{Y} - \hat{b}_1 \bar{X} = 0 \quad (3)$$

Sub. this into (2) gives

$$\hat{b}_1 \sum X_i^2 - \sum X_i Y_i + (\bar{Y} - \hat{b}_1 \bar{X}) \sum X_i = 0$$

$$\hat{b}_1 \sum X_i^2 - \sum X_i Y_i + (\bar{Y} - \hat{b}_1 \bar{X}) n \bar{X} = 0$$

$$\hat{b}_1 (\sum X_i^2 - n \bar{X}^2) = \sum X_i Y_i - n \bar{X} \bar{Y}$$

Dividing both sides by $1/N$

$$\hat{b}_1 \left(\frac{1}{N} \sum X_i^2 - \bar{X}^2 \right) = \frac{1}{N} \sum X_i Y_i - \bar{X} \bar{Y}$$

Which gives the formula to calculate the OLS estimate of the slope

$$\hat{b}_1 \text{Var}(X) = \text{Cov}(X, Y)$$

$$\hat{b}_1 = \frac{\text{Cov}(X, Y)}{\text{Var}(X)} \quad (4)$$

(4) has a nice, clear intuitive meaning about the influence of the variable X on the size of the slope, since it shows that

- a) the greater the covariance between X and Y
- b) the smaller the variance of X

the larger the (absolute value of the) OLS estimate of \hat{b}_1

b) Multiple Regression Analysis

In most cases unlikely can explain all of behaviour in the dependent variable by a single explanatory variable. Most problems require 2 or more right hand side variables to capture behaviour adequately.

Consider model extended to k variables for i=1, 2 ... N individuals

$$Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \dots + \hat{\beta}_k X_{ki} + u_i$$

Useful to write down model in a more compact matrix notation

$$\underset{(N \times 1)}{\mathbf{y}} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} \quad \underset{N \times k}{\mathbf{X}} = \begin{bmatrix} 1 & X_{11} & \dots & X_{k1} \\ 1 & X_{12} & \dots & X_{k2} \\ \vdots & \vdots & \dots & \vdots \\ 1 & X_{1N} & \dots & X_{kN} \end{bmatrix} \quad \underset{(N \times 1)}{\mathbf{u}} = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_N \end{bmatrix} \quad \underset{(k \times 1)}{\boldsymbol{\beta}} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix}$$

Minimising the sum of squared residuals now implies

$$\begin{aligned} \text{Min}_{\boldsymbol{\beta}} \hat{\mathbf{u}}' \hat{\mathbf{u}} &= \begin{bmatrix} \hat{u}_1 & \hat{u}_2 & \dots & \hat{u}_n \end{bmatrix} \begin{bmatrix} \hat{u}_1 \\ \hat{u}_2 \\ \vdots \\ \hat{u}_N \end{bmatrix} = \hat{u}_1^2 + \hat{u}_2^2 + \dots + \hat{u}_N^2 \\ &= (\mathbf{y} - \mathbf{X} \hat{\boldsymbol{\beta}})' (\mathbf{y} - \mathbf{X} \hat{\boldsymbol{\beta}}) = \mathbf{y}' \mathbf{y} - \hat{\boldsymbol{\beta}}' \mathbf{X}' \mathbf{y} - \mathbf{y}' \mathbf{X} \hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\beta}}' \mathbf{X}' \mathbf{X} \hat{\boldsymbol{\beta}} \end{aligned}$$

Since all terms are scalars (1x1) can add

$$= \mathbf{y}' \mathbf{y} - 2 \hat{\boldsymbol{\beta}}' \mathbf{X}' \mathbf{y} + \hat{\boldsymbol{\beta}}' \mathbf{X}' \mathbf{X} \hat{\boldsymbol{\beta}}$$

F.O.C. minimum

$$\frac{\partial \hat{\mathbf{u}}' \hat{\mathbf{u}}}{\partial \hat{\boldsymbol{\beta}}} = -2 \mathbf{X}' \mathbf{y} + 2 \mathbf{X}' \mathbf{X} \hat{\boldsymbol{\beta}} = 0$$

which gives k normal equations $\mathbf{X}' \mathbf{X} \hat{\boldsymbol{\beta}} = \mathbf{X}' \mathbf{y}$

and the k variable OLS solution $\hat{\boldsymbol{\beta}} = (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}' \mathbf{y}$

Algebraic Aspects of OLS

$$\text{Given } X'X\hat{\beta} = X'y$$

$$\text{Re-write as } X'X\hat{\beta} - X'y = 0$$

$$\Rightarrow -X'(y - X\hat{\beta}) = 0$$

$$\text{so } X'u = 0$$

It follows that

$$1) \text{ OLS residuals add to zero } \sum_{i=1}^N \hat{u}_i = 0$$

2) Mean of OLS residuals is zero

3) Regression passes through the point of means in K dimensional space

$$\hat{u} = \bar{y} - \bar{X}\hat{\beta} = 0$$

4) Each regressor variable is uncorrelated with the OLS residuals

$$X'u = 0 \Leftrightarrow \text{Cov}(X, u) = 0$$

5) The set of predicted values are uncorrelated with the residuals

$$\hat{y}'u = 0$$

6) Mean of estimated values equals the mean of actual values

$$\bar{\hat{y}} = \bar{y}$$

Partialing Out in Multiple Regression

OLS estimates can be interpreted as partial derivatives ie $\hat{\beta}_k$ is the effect of a unit change in the level of variable X_k holding all other variables constant

Given

$$y = X \hat{\beta} + u$$

Consider partitioning the X matrix into $X = [x_1 : X_2]$

ie the $N \times 1$ vector of observations on a single variable (x_1)

and

the $N \times (k-1)$ matrix of observations on the other $k-1$ right-hand side variables (including the constant)

so

$$y = \begin{bmatrix} x_1 & X_2 \end{bmatrix} \begin{bmatrix} \hat{b}_1 \\ \hat{\beta}_1 \end{bmatrix} + u$$

Given OLS normal equations $X' X \hat{\beta} = X' y$

$$\begin{bmatrix} x_1' \\ X_2' \end{bmatrix} \begin{bmatrix} x_1 & X_2 \end{bmatrix} \begin{bmatrix} \hat{b}_1 \\ \hat{\beta}_1 \end{bmatrix} = \begin{bmatrix} x_1' \\ X_2' \end{bmatrix} y$$

so

$$\begin{bmatrix} x_1' x_1 & x_1' X_2 \\ X_2' x_1 & X_2' X_2 \end{bmatrix} \begin{bmatrix} \hat{b}_1 \\ \hat{\beta}_1 \end{bmatrix} = \begin{bmatrix} x_1' y \\ X_2' y \end{bmatrix}$$

1st row can be written

$$(x_1' x_1)^{-1} \hat{b}_1 + (x_1' X_2) \hat{\beta}_1 = x_1' y$$

so

$$\hat{b}_1 = (x_1' x_1)^{-1} x_1' y - (x_1' x_1)^{-1} (x_1' X_2) \hat{\beta}_1$$

Hence OLS estimate of coefficient b_1 in multiple regression is the same as that obtained in a simple regression together with a correction factor that takes account of the association between x_1 and the other variables

If $X_1' X_2 = 0$ variables said to be orthogonal only then is multiple regression estimate is identical to that in a simple regression

Frisch-Waugh Theorem

-useful for idea that multiple regression coefficients “partial out” effects of other variables

Let $M_2 = I - X_2(X_2'X_2)^{-1}X_2'$

["residual maker" matrix since when multiplied by y gives the residuals from an OLS regression of y on X₂]

$$\hat{u}_2 = y - X_2 \hat{\beta} = y - X_2(X_2'X_2)^{-1}X_2'y = M_2y$$

M₂ is also an "idempotent" matrix such that M₂=M₂'M₂

Given $y = x_1 \hat{b}_1 + X_2 \hat{\beta}_2 + \hat{u}$

$$M_2y = M_2x_1 \hat{b}_1 + M_2X_2 \hat{\beta}_2 + M_2\hat{u}$$

$$M_2y = M_2x_1 \hat{b}_1 + \hat{u}$$

pre-multiply by x₁'

$$x_1'M_2y = x_1'M_2x_1 \hat{b}_1$$

so

$$\hat{b}_1 = (x_1'M_2'M_2x_1)^{-1}x_1'M_2'M_2y$$

which since

M₂y is vector of residuals when y is regressed on X₂

M₂x₁ is vector of residuals when x₁ is regressed on X₂

Says that the any one multiple regression coefficient can also be obtained by netting out the effect of the other variables on both the dependent and independent variable of interest

Regression Diagnostics and Influential Data Points

In data sets with small number of observations, useful to ascertain whether any one individual observation is particularly influential (results could change significantly if observation were removed or added).

Consider “Hat” matrix

$$H = X(X'X)^{-1}X'$$

and

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

Gives $n \times 1$ vector of OLS estimates of predicted y values

$$\text{So any one predicted value } \hat{y}_i = H_i y$$

(where H_i is the i^{th} row of X)

is a weighted average of all the elements of the y vector

$$\hat{y}_i = H_{i1}y_1 + H_{i2}y_2 + \dots + H_{iN}y_N$$

where the j^{th} weight, H_{ij} reflects the contribution of y_j to the predicted value which in turn depends on the relative size of each observation on the X variable. Depending on their observed value, certain observations on X can have a larger “weight” and affect the OLS estimates more than others

[Similarly since

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

then OLS estimates are also a weighted average of the elements of the y vector]

Let the scalar value

$$h_i = x_i(X'X)^{-1}x_i' \quad (1)$$

where x_i is the i^{th} row of X

be the i^{th} element on the main diagonal of H . Said to be the "leverage" of the i^{th} observation

In 2 variable model can show (Besley, Kuh, Welsch (1980)) that

$$h_i \approx \frac{1}{N} + \frac{(x_i - \bar{X})^2}{\sum_{j=1}^N (x_j - \bar{X})^2} \quad (2)$$

$0 \leq h_i \leq 1$
(1 is high leverage)

which depends on the deviation of the i^{th} observation on the X variable relative to the average deviation

(In the k variable model equivalent above, h_i effectively measures the distance away from K means)

Can also show (Davidson & MacKinnon ch. 2) that the difference between OLS estimates from a regression with the influential observation, $\hat{\beta}$, and that without, $\hat{\beta}_i$, is given by

$$\hat{\beta}_i - \hat{\beta} = \frac{-1}{1 - h_i} (X'X)^{-1} X' u_i$$

(where \hat{u}_i is the vector of OLS residuals when the influential observation is *excluded* and X is the matrix of observations *including* the influential observation)

Any observation with a large leverage will pull the regression line toward it
- though (2) indicates that influence is reduced as sample size, N , increases

Note also that observations with a high degree of leverage will tend to have a smaller residual.

Observations with large residuals are often called outliers

Useful therefore to study both leverage and the size of the residuals from individual observations in small data sets

Often use standardised residuals to do this

$$r_i = \frac{\hat{u}_i}{s\sqrt{1-h_i}}$$

where s = standard error of regression equation = $\sqrt{\text{RSS}/(N-k)}$

(normalised by its standard error becomes scale invariant)

N.B. Studentised residual

$$\tilde{u}_i = \frac{\hat{u}_i}{s_i\sqrt{1-h_i}}$$

where s_i = standard error of regression equation = $\sqrt{\text{RSS}/(N-k)}$

Tests for Influential Observations

1) Inspection

2) DFITS test

- summarises contribution of both

$$= r_i \sqrt{\frac{h_i}{1-h_i}}$$

can show that $DFITS > 2\sqrt{(k/N)}$ is worth investigating

3) Cook's Distance

$$= \frac{1}{k} \frac{s_i^2}{s^2} DFITS^2$$

and if $> 4/N$ the observation should be examined

Ultimately even if observation gives cause for concern, need a good reason as to why should drop it from the data set (eg measured with error or a "one-off" shock).

Alternative methods of dealing with outliers include quantile regression estimation

Goodness of Fit

Useful to have summary measure of how well the OLS regression line fits the data
Given

$$y = X\hat{\beta} + \hat{u} = \hat{y} + \hat{u}$$

$$y'y = (X\hat{\beta} + \hat{u})'(X\hat{\beta} + \hat{u})$$

$$y'y = \hat{\beta}'X'X\hat{\beta} + \hat{u}'\hat{u}$$

Since $y'y \neq \sum_i (y_i - \bar{y})^2$

Need to subtract $N\bar{y}^2$ from both sides

$$(y'y - N\bar{y}^2) = (\hat{\beta}'X'X\hat{\beta} - N\bar{y}^2) + \hat{u}'\hat{u}$$

Total sum of Squares	Explained sum of squares	+ residual sum of squares
----------------------	--------------------------	---------------------------

Hence measure of how well regression fits the data is given by

$$R^2 \text{ (the coefficient of determination) } = \text{ESS/TSS} = 1 - (\text{RSS/TSS})$$

- the proportion of the total variation in y accounted for by the variation in the regressors

$$0 \leq R^2 \leq 1$$

Major problem with using r^2 is that will never fall when add new explanatory variables to the model

Use instead the Adjusted R^2 ,
$$\bar{R}^2 = 1 - \frac{(N-1)}{(N-k)}(1 - R^2)$$

Whether this rises or falls depends on whether variable added to model has a t ratio greater than one

There are other goodness of fit criteria that impose different weights to additional numbers of regressors

1. Schwarz $\text{Log}_e(\text{RSS}/N) + k/N \text{log}_e(N)$

2. Akaike $\text{Log}_e(\text{RSS}/N) + 2k/N$

these penalties are larger than those imposed in the adjusted R^2 and so will favour simpler models

Assumptions Underlying OLS

In order to assess the accuracy or the precision of OLS estimates need to make assumptions about the statistical process generating the observations in the dataset.

1. regression model is linear in parameters
2. X values are fixed in repeated sampling
3. Mean value of (true, unobserved) residuals is zero
4. No covariance between residuals and independent variables
5. Equal variance of individual residual terms (homoskedasticity)
6. No correlation between individual residual terms (no autocorrelation)

When moving from 2 to k variable model need to make 1 additional assumption

7. No exact colinearity between X variables

Statistical Properties of OLS Estimators

$$\text{Given } \hat{\beta} = (X'X)^{-1}X'y = (X'X)^{-1}X'(X\beta + u) = \beta + (X'X)^{-1}X'u$$

So

$$E(\hat{\beta}) = \beta + (X'X)^{-1}X'E(u)$$

OLS estimators are unbiased

$$\begin{aligned} \text{Var}(\hat{\beta}) &= E\left[(\hat{\beta} - E(\hat{\beta}))(\hat{\beta} - E(\hat{\beta}))'\right] \\ \text{Var}(\hat{\beta}) &= E\left[(\hat{\beta} - \beta)(\hat{\beta} - \beta)'\right] = E\left[(X'X)^{-1}X'uu'X(X'X)^{-1}\right] \\ &= (X'X)^{-1}X'E(uu')X(X'X)^{-1} \\ &= (X'X)^{-1}X'\sigma^2I X(X'X)^{-1} \\ &= \sigma^2(X'X)^{-1} \end{aligned}$$

which is a $k \times k$ matrix with variances on the main diagonal and covariances on the off diagonal

Gauss-Markov Theorem

Consider another linear unbiased estimator $\tilde{\beta} = Cy$

If $\tilde{\beta}$ is unbiased then $E(\tilde{\beta}) = E(Cy) = E[CX\beta + Cu] = \beta$

Hence $CX=I$ and $\tilde{\beta} = \beta + Cu$

$$\begin{aligned} \text{So } \text{Var}(\tilde{\beta}) &= E\left[(\tilde{\beta} - \beta)(\tilde{\beta} - \beta)'\right] \\ &= E[Cuu'C'] \\ &= \sigma^2CC' \end{aligned}$$

Let D be the difference between the OLS and alternative estimated explanatory component ie

$$D = C - (X'X)^{-1}X'$$

So

$$\text{Var}(\tilde{\beta}) = \sigma^2 \left[(D + (X'X)^{-1}X')(D + (X'X)^{-1}X')' \right]$$

Since $CX = I = DX + (X'X)^{-1}X'X$ the $DX = 0$

Cross product terms vanish and

$$\text{Var}(\tilde{\beta}) = \sigma^2 DD' + \sigma^2 (X'X)^{-1} = \sigma^2 DD' + \text{Var}(\hat{\beta})_{\text{OLS}}$$

ie variance of alternative estimator equals that of OLS plus a non-negative definite matrix (see problem set 0)

Hence OLS estimate has minimum variance property (BLUE – Best Linear Unbiased Estimator). Main reason for widespread use of OLS, will always provide estimators with smaller standard errors