Lecture 2. Algebraic Aspects of OLS

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

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

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

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

It follows that

1) 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

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

4) Each regressor variable is uncorrelated with the OLS residuals

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

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

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

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

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

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

ie the N_x1 vector of observations on a single variable (x_1)
and
the $N_x(k-1)$ matrix of observations on the other k-1 right-hand side variables (including the constant)

so $$y = [x_1 \quad X_2]\begin{bmatrix}\hat{b}_1 \\ \hat{\beta}_2\end{bmatrix} + \hat{u}$$

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

$$\begin{bmatrix}x_1' \\ X_2'\end{bmatrix}[x_1 \quad X_2]\begin{bmatrix}\hat{b}_1 \\ \hat{\beta}_2\end{bmatrix} = \begin{bmatrix}x_1' \\ X_2'\end{bmatrix}y$$

so

$$\begin{bmatrix}x_1'x_1 & x_1'X_2 \\ X_2'x_1 & X_2'X_2\end{bmatrix}\begin{bmatrix}\hat{b}_1 \\ \hat{\beta}_2\end{bmatrix} = \begin{bmatrix}x_1'y \\ X_2'y\end{bmatrix}$$

1st row can be written

$$(x_1'x_1)\ \hat{b}_1 + (x_1'X_2)\hat{\beta}_2 = x_1'y$$

so

$$\hat{b}_1 = (x_1'x_1)^{-1}x_1'y - (x_1'x_1)^{-1}(x_1'X_2)\hat{\beta}_2$$

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 X_2]

$$\hat{u}_2 = y - X_2\hat{\beta} = y - X_2(X_2'X_2)^{-1}X_2'y = M_2 y$$

M_2 is also an “idempotent” matrix such that $M_2=M_2'M_2$

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

$$M_2 y = M_2 x_1\hat{b}_1 + M_2 X_2\hat{\beta}_2 + M_2\hat{u}$$

$$M_2 y = M_2 x_1\hat{b}_1 + \hat{u}$$

pre-multiply by x_1'

$$x_1' M_2 y = x_1' M_2 x_1\hat{b}_1$$

so

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

which since

M_2y is vector of residuals when y is regressed on X_2
M_2x_1 is vector of residuals when x_1 is regressed on X_2

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

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

Gives n_x1 vector of OLS estimates of predicted y values

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

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

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

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

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

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

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

Let the scalar value

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

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

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

$0 <= h_i <= 1$
(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

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

(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 = 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 = 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√(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