Algebra of the linear regression model

Giuseppe Ragusa

12 Mar 2016

The model in matrix form

$$y = X\beta + u$$

$$\underset{(n\times 1)}{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \quad \underset{(n\times k)}{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & & x_{2k} \\ \vdots & & & & \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix} \quad \underset{(n\times 1)}{u} = \begin{pmatrix} u1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix}$$

Algebraic Aspects of OLS

The least squares problem

$$\arg\min_{\beta}(Y - X\beta)'(Y - X\beta)$$

The first order conditions

$$-X'Y + X'X\beta = -X'u = 0$$

The solution

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

Basic Algebraic Aspects of OLS

- *If there is a constant, the least squares residuls sum to zero*

$$X'(Y - X\hat{\beta}) = 0 \implies \sum_{i=1}^{n}(Y - X\hat{\beta}) = 0$$

- $\sum_{i=1}^{n}(Y - X\hat{\beta}) = 0$ implies that

$$\bar{Y} = \bar{X}\hat{\beta},$$

  that is, *the regression hyperplane passes through the point of means of the data;*

- Notice that, for $i = (1, 1, \ldots, 1)$

$$i'\hat{Y} = i'(X\hat{\beta}) = \bar{X}\hat{\beta},$$

  thus

$$\bar{\hat{Y}} = \bar{Y},$$

  the mean of fitted values from the regression equals the mean of the actual data.

Projection

$$\hat{u} = Y - X\hat{\beta} = Y - X(X'X)^{-1}X'Y = (I_n - X(X'X)^{-1}X')Y = MY$$

- The $n \times n$ matrix M is foundamental in regression analysis
- M has some important properties
  - M is symmetric ($M = M'$)
  - M is idempotent ($M = MM$)
  - M is orthogonal $MX = 0$
- $\hat{Y} = Y - \hat{u} = Y - MY = (I - M)Y = PY$
  - P is a projection matrix
  - $PM = MP = 0$
  - $PX = X$

Partitioned Regression and Partial Regression

$$Y = X_1\beta_1 + X_2\beta_2 + u$$

The first order conditions for β_1 and β_2 are

$$X_1'X_1\beta_1 + X_1'X_2\beta_2 = X_1'y$$
$$X_2'X_1\beta_1 + X_2'X_2\beta_2 = X_2'y$$

It follows

$$\hat{\beta}_1 = (X_1'X_1)^{-1}X_1'y + (X_1'X_1)^{-1}X_1'X_2\beta_2$$
$$\hat{\beta}_2 = (X_2'X_2)^{-1}X_2'y + (X_2'X_2)^{-1}X_2'X_1\beta_1$$

Partitioned Regression and Partial Regression, ctd.

Theorem

In the multiple linear regression of Y on two sets of variables, X_1 and X_2, if X_1 and X_2 are orthogonal the coefficient $\hat{\beta}_1$ and $\hat{\beta}_2$ can be obtained by regressing Y on X_1 and Y on X_2 separately.

Proof.

If X_1 and X_2 are orthogonal, $X_1'X_2 = X_2'X_1 = 0$. Thus the result follows from the equation for $\hat{\beta}_1$ and $\hat{\beta}_2$. $\square$

Partitioned Regression and Partial Regression, ctd.

Theorem (Frisch-Waugh-Lovell)

In the multiple linear regression of Y on two sets of variables, X_1 and X_2, the subvector $\hat{\beta}_2$ is the set of coefficient obtained when the residuals from a regression of Y on X_1 alone are regressed on the set of residuals obtained when each column of X_2 is regressed on X_1.

Proof.

Substituting the expression for $\hat{\beta}_1$ into the expression for β_2

$$\begin{aligned}\hat{\beta}_2 &= (X_2'X_2)^{-1}X_2'y + (X_2'X_2)^{-1}X_2'X_1(X_1'X_1)^{-1}X_1'y + (X_1'X_1)^{-1}X_1'X_2\beta_2 \\ &= (X_2'M_1X_2)^{-1}X_2'M_1Y\end{aligned}$$

where $M_1 = (I - X_1'(X_1'X_1)^{-1}X_1')$. The result follows from the idempotency of M_1 and the fact that $X_2'M_1$ and M_1Y are the residuals of a regression of X_2 on X_1 and of Y on X_1.

$\square$

Partial correlation

Consider

$$Y = X_1\beta + \gamma z + u$$

where z is a n vector.
The partial correlation coefficient between Y and z is

$$r^*_{Yz} = \frac{z_*' Y_*}{z_*' z_*}(Y_*' Y_*)$$

where

1. $Y_* = M_1 Y$
2. $z_* = M_1 z$

Partial correlation

The partial correlation can be obtained directly from the multiple regression of Y on X_1 and z

$$r^*_{Yz} = \frac{t_z^2}{t_z^2 + n - (k+1)}$$

where t_z is the t-ratio for testing the null hypothesis $H_0 : \gamma = 0$ calculated on the full regression of Y on X_1 and z.