Matrix algebra and the linear regression model

Susan Thomas
IGIDR, Bombay

25 September 2008

Model $Y_i = \beta_1 X_i + \epsilon_i$

- Consider a two variable, gaussian distribution model *without an intercept*:
  - Independence: Y_i, X_i are independent across i
  - Normality conditional on X_i: $Y_i \sim N[\beta_1 X_i, \sigma^2]$
  - X_i is exogenous
- $l = -\frac{N}{2} \log{(2\pi\sigma^2)} - \frac{1}{2\sigma^2} \sum_{i=1}^{N} (Y_i - \beta_1 X_i)^2$

MLE $\hat{\beta}_1$

- $\frac{\partial l}{\partial \beta_1} = \sum_{i=1}^{N}(Y_i - \beta_1 X_i)X_i = 0$
- $\hat{\beta}_1 = \sum_{i=1}^{N}(Y_i X_i) / \sum_{i=1}^{N}(X_i X_i)$

MLE $\hat{\sigma}^2$

- $\frac{\partial l}{\partial \sigma^2} = -\frac{N}{(2\sigma^2)} + \frac{1}{2\sigma^4}\sum_{i=1}^{N}(Y_i - \hat{\beta}_1 X_i)^2 = 0$

$$\begin{aligned}
\hat{\sigma}^2 &= \frac{1}{N}\sum_{i=1}^{N}(Y_i - \hat{\beta}_1 X_i)^2 \\
&= \frac{1}{N}\sum_{i=1}^{N}(Y_i^2 - 2\hat{\beta}_1 X_i Y_i - \hat{\beta}^2 X_i^2) \\
&= \frac{1}{N}\sum_{i=1}^{N}(Y_i^2 - 2\frac{\sum_{i=1}^{N}(Y_i X_i)}{\sum_{i=1}^{N}(X_i)^2} X_i Y_i - \frac{\sum_{i=1}^{N}(Y_i X_i)^2}{\sum_{i=1}^{N}(X_i)^2} X_i^2) \\
&= \frac{1}{N}(\sum Y_i^2 - \frac{\sum_{i=1}^{N}(Y_i X_i)\sum_{i=1}^{N}(Y_i X_i)}{\sum_{i=1}^{N} X_i^2}) \\
&= \frac{1}{N}(\sum Y_i^2 - \sum_{i=1}^{N}(Y_i X_i)(\sum_{i=1}^{N} X_i^2)^{-1}\sum_{i=1}^{N}(Y_i X_i))
\end{aligned}$$

Is hard work!

Matrix notation gives simplicity and generalisation

Matrix notation for $Y_i = \beta_1 X_i + \epsilon_i$

- Y_i, X_i are N-dimensional vectors:

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \dots \\ Y_N \end{pmatrix}, \qquad X = \begin{pmatrix} X_1 \\ X_2 \\ \dots \\ X_N \end{pmatrix}$$

- β is a 1-dimensional vector in this problem.
- $X'Y = (X_1, \dots, X_N) \begin{pmatrix} Y_1 \\ Y_2 \\ \dots \\ Y_N \end{pmatrix} = \sum_{i=1}^{N} X_i Y_i$
- $X'X = (X_1, \dots, X_N) \begin{pmatrix} X_1 \\ X_2 \\ \dots \\ X_N \end{pmatrix} = \sum_{i=1}^{N} X_i X_i$

The Data Matrix

- The data is a matrix $D = (Y, X)$ such that

$$D = (Y, X) = \begin{pmatrix} Y_1 & X_1 \\ Y_2 & X_2 \\ & \cdots \\ Y_N & X_N \end{pmatrix}$$

- Then, a convenient matrix is $D'D =$:

$$(Y, X)'(Y, X) = \begin{pmatrix} Y' \\ X' \end{pmatrix} (Y, X) = \begin{pmatrix} Y'Y & Y'X \\ X'Y & X'X \end{pmatrix}$$

- Features of $D'D$: Diagonal are variances and positive. Symmetric about the diagonal.

Matrix notation in the MLE framework

- The model for y_i without an intercept is $y_i = \beta_1 x_i + \epsilon_i$
- Matrix form:

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \dots \\ Y_N \end{pmatrix} = \begin{pmatrix} X_1 \\ X_2 \\ \dots \\ X_N \end{pmatrix} \beta + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \dots \\ \epsilon_N \end{pmatrix}$$

  Note: Dimensionality of $Y = X = \epsilon = N \times 1$.
- $Y = X\beta + \epsilon$

Matrix notation in the MLE framework

- $l = l_{Y|X}(\beta, \sigma^2) = -\frac{N}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(Y - X\beta)'(Y - X\beta)$
- Maximising l is the same as minimising SSE $= (Y - X\beta)'(Y - X\beta)$ (with $(Y - X\beta) = \epsilon$).
- As before, the maximum is obtained by setting $\partial l / \partial \beta = 0$.
- We can use matrix calculus to find out that:

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

- Similarly, we find that:

$$\hat{\sigma}^2 = \frac{1}{N}(Y'Y - Y'X(X'X)^{-1}X'Y)$$

Note: This notation is convenient, because more exogenous variables will have the same solution form in this matrix notation.

Getting the MLE RSS from the matrix form

- Now if $\hat{\beta} = (X'X)^{-1}(X'Y)$ then $\hat{\beta}' = (Y'X)(X'X)^{-1}$ and

$$\begin{pmatrix} 1 & -Y'X(X'X)^{-1} \\ 0 & 1 \end{pmatrix} \begin{pmatrix} Y'Y & Y'X \\ X'Y & X'X \end{pmatrix} \begin{pmatrix} 1 & 0 \\ -(X'X)^{-1}X'Y & 1 \end{pmatrix}$$

$$= \begin{pmatrix} Y'Y - Y'X(X'X)^{-1}X'Y & 0 \\ 0 & X'X \end{pmatrix}$$

- Which is:

$$\begin{pmatrix} 1 & -\hat{\beta}' \\ 0 & 1 \end{pmatrix} D'D \begin{pmatrix} 1 & 0 \\ -\hat{\beta} & 1 \end{pmatrix} = \begin{pmatrix} N\hat{\sigma}^2 & 0 \\ 0 & X'X \end{pmatrix}$$

Recap: on *se* of $\hat{\beta}$

- Recall for $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$
- $E(\hat{\beta}_1) = \beta$
- $\text{var}(\hat{\beta}_1) = \sigma_\epsilon^2 / (\sum_{i=1}^{N} (X_i - \bar{X})^2$
- Therefore, $\begin{pmatrix} 1 & -\hat{\beta}' \\ 0 & 1 \end{pmatrix} D'D \begin{pmatrix} 1 & 0 \\ -\hat{\beta} & 1 \end{pmatrix}$ contain the elements to calculate the variance of the MLE $\hat{\beta}$.

Rewriting the model SSE in matrix notation

- SSE = $(Y - X\beta)'(Y - X\beta)$
- With a sample, we will have estimated $\hat{\beta}$.
  This gives us an estimated $\hat{Y} = X\hat{\beta}$.
- SSE can be rewritten using the estimate, $\hat{Y}$ as follows:

$$\begin{aligned} \epsilon &= Y - X\beta = Y - X\hat{\beta} + X\hat{\beta} - X\beta \\ &= (Y - X\hat{\beta}) + X(\hat{\beta} - \beta) \\ \text{Estimated residual, } \hat{\epsilon} &= Y - X\hat{\beta} \\ \text{SSE} = \epsilon'\epsilon &= \hat{\epsilon}\hat{\epsilon}' + (\hat{\beta} - \beta)'X'X(\hat{\beta} - \beta) \end{aligned}$$

- $\hat{\epsilon}\hat{\epsilon}'$ is the estimated SSE.
- $\hat{\beta} - \beta$ is the error in the estimation of β.
- The SSE will be a minimum when $\hat{\beta} - \beta = 0$.
  Ie, when the estimation error is zero and $\hat{\beta} = \beta$

The two variable regression model with intercept

Matrix notation for $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$

- $Y = X\beta + \epsilon$

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \ldots \\ Y_N \end{pmatrix} = \begin{pmatrix} 1 & X_1 \\ 1 & X_2 \\ \ldots & \\ 1 & X_N \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \ldots \\ \epsilon_N \end{pmatrix}$$

which gives

$$\begin{aligned} Y_1 &= \beta_0 + \beta_1 X_1 + \epsilon_1 \\ Y_2 &= \beta_0 + \beta_1 X_2 + \epsilon_2 \\ \ldots Y_N &= \beta_0 + \beta_1 X_N + \epsilon_N \end{aligned}$$

- Since the form of the model remains the same, $Y = X\beta + \epsilon$, we can use the same form for the esimate $\hat{\beta} = (X'X)^{-1}X'Y$

Solution $\hat{\beta}$ for the two-variable problem

- $X'X$ is: $\begin{pmatrix} \sum_{i=1}^{N} 1^2 & \sum_{i=1}^{N} X_i \\ \sum_{i=1}^{N} X_i & \sum_{i=1}^{N} X_i^2 \end{pmatrix}$
- $X'Y$ is: $\begin{pmatrix} \sum_{i=1}^{N} Y_i \\ \sum_{i=1}^{N} X_i Y_i \end{pmatrix}$
- $\beta = (X'X)^{-1}X'Y \rightarrow (X'X)\beta = X'Y$

$$\begin{pmatrix} N & \sum_{i=1}^{N} X_i \\ \sum_{i=1}^{N} X_i & \sum_{i=1}^{N} X_i^2 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{N} Y_i \\ \sum_{i=1}^{N} X_i Y_i \end{pmatrix}$$

- This gives the same first derivative equations as before:

$$N\beta_0 + \beta_1 \sum_{i=1}^{N} X_i = \sum_{i=1}^{N} Y_i$$

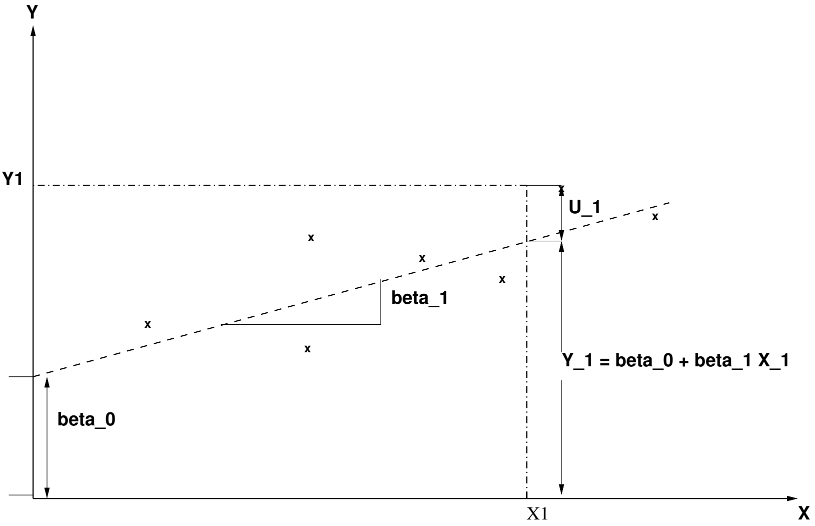
$$\beta_0 \sum_{i=1}^{N} X_i + \beta_1 \sum_{i=1}^{N} X_i^2 = \sum_{i=1}^{N} Y_i X_i$$

The linear regression model

$Y_i = \beta_0 + X\beta + u_i$

- Y_i dependent variable
- X_i independent variable
- u_i error, random variable, i.i.d.
- $(\beta_0, \beta_1, \ldots, \beta_{X_i})$ population parameters.
- i subscript, denoting data points
- Simplest: Only one independent variable, x_1
  1. $\beta_0 + \beta_1 X_i$ is the deterministic part of the model.
     It is the conditional mean of Y_i i.e. $E(Y_i|X_i) = \beta_0 + \beta_1 X_i$
     when $E(u_i|X_i) = 0$
  2. Linear: Linear in parameters, Not variables.
  3. $u_i \rightarrow$ difference between Y_i and $(\beta_0 + \beta_1 X_i)$
     $Y_i - \beta_0 - \beta_1 X_i = u_i$

Implication of the "population" relation beween Y, X



Interpreting the role of u_i

u_i is included in the regression to accomodate at least four types of effects:

1. Omitted variables
2. Non-linearities in X
3. Measurement error (in Y, X)
4. Randomness of behaviour/ effects.

Interpretation of β_0, β_1

- Interpretation of β_1: marginal effect of X_i on Y_i
- Interpretation of β_0: less simple.
  It could contain the effect of all the omitted variables and other effects.
- Model error: $Y_i - \beta_0 + \beta_1 X_i = u_i$

From population to sample

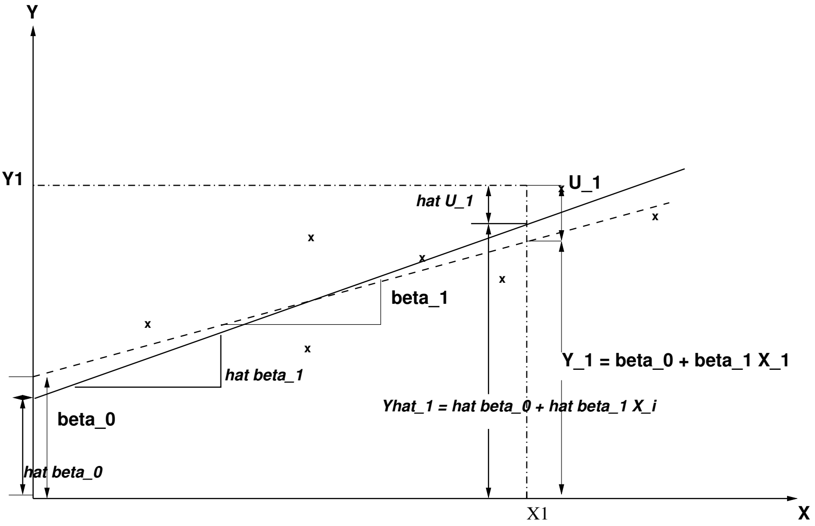
- We estimate $\hat{\beta}_0, \hat{\beta}_1$.
- We use this to get $\hat{Y}_i$.
  This is the *estimated value* of Y_i.
- The estimated error is $\hat{u}_i = Y_i - \hat{Y}_i$.
- Both "population regression" and the "sample regression" have errors.

$$u_i = Y_i - \beta_0 - \beta_1 X_i$$

$$\hat{u}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i$$

- Objective of regression approach is the get the "best" $\hat{\beta}_1$.
- Given the model, the best that we can get is $\hat{\beta}_1 = \beta_1$.
  This will give us $\text{SSE}_{\hat{u}_i}$ will be the same as SSE_{u_i}.

Implication of the regression beween Y, X



Estimation by Ordinary Least Squares

- Method of OLS: choose $(\hat{\beta}_0, \hat{\beta}_1)$ to minimize the sum of squared errors (SSE).

$$\begin{aligned} \hat{u}_i &= Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i \\ \hat{u}_{i^2} &= (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2 \end{aligned}$$

- $ESS \equiv \sum \hat{u}_i^2 = \sum (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2$

$$\text{Minimize } \hat{\beta}_0, \hat{\beta}_1, X_i \quad \sum \hat{u}_i^2 = \sum (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2$$

The rationale of OLS

Why calculate the SSE?

1. Strips the direction of errors
2. Penalises large errors.

$$\begin{array}{ccc} (-1, 2, +1, -2) & \longrightarrow & (-1, -1, -1, 3) \\ ESS = 10 & & ESS = 12 \end{array}$$

The OLS solution

- This approach gives us what is called the *normal equations*:

$$\begin{aligned} ESS &= \sum (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2 \\ \frac{\partial ESS}{\partial \hat{\beta}_0} &= \sum 2(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)(-1) \\ \frac{\partial ESS}{\partial \hat{\beta}_1} &= \sum 2(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)(-X_i) \end{aligned}$$

- Set these to zero (just as in the case of the MLE):

$$\begin{aligned} \sum 2(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)(-1) &= 0 \\ \sum 2(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)(-X_i) &= 0 \end{aligned}$$

The OLS solution

- Two equations in two unknowns($\hat{\beta}_0, \hat{\beta}_1$). Solve them

$$\begin{aligned}
\rightarrow \quad & \sum (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = 0 \\
\Rightarrow \quad & \sum Y_i - N\hat{\beta}_0 - \hat{\beta}_1 \sum X_i) = 0 \\
\Rightarrow \quad & N\hat{\beta}_0 = \sum Y_i - \hat{\beta}_1 \sum X_i \\
\Rightarrow \quad & \hat{\beta}_0 = \frac{1}{N} \sum Y_i - \hat{\beta}_1 \frac{1}{N} \sum X_i
\end{aligned}$$

$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$
Sample regression line passes through mean.

The OLS solution

$$\Rightarrow \qquad \sum \left[Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i \right] [X_i] = 0$$

$$\text{or,} \quad \sum Y_i X_i - \hat{\beta}_0 \sum X_i - \hat{\beta}_1 \sum X_i^2 = 0$$

$$\text{or,} \quad \sum Y_i X_i = \hat{\beta}_0 \sum X_i + \hat{\beta}_1 \sum X_i^2$$

$$\text{or,} \quad \sum Y_i X_i = \left[\frac{1}{N} \sum Y_i - \hat{\beta}_1 \frac{1}{N} \sum X_i \right] \sum X_i + \hat{\beta}_1 \sum X_i^2$$

$$\text{or,} \quad \sum Y_i X_i = \frac{1}{N} \sum Y_i \sum X_i - \hat{\beta}_1 \frac{1}{N} \sum X_i \sum X_i + \hat{\beta}_1 \sum X_i^2$$

$$\text{or,} \quad \sum Y_i X_i = \frac{1}{N} \sum Y_i \sum X_i + \hat{\beta}_1 \left[\sum X_i^2 - \frac{1}{N} (\sum X_i)^2 \right]$$

$$\hat{\beta}_1 = \frac{\sum Y_i X_i - \frac{1}{N} \sum Y_i \sum X_i}{\sum X_i^2 - \frac{1}{N} (\sum X_i)^2}$$

The OLS solution

$$\begin{aligned}\sum (X_i - \bar{X})(Y_i - \bar{Y}) &= \sum (X_i Y_i - X_i \bar{Y} - \bar{X} Y_i + \bar{X}\bar{Y}) \\ &= \sum X_i Y_i - \bar{Y} \sum X_i - \bar{X} \sum Y_i + \sum \bar{X}\bar{Y} \\ &= \sum X_i Y_i - N\bar{X}\bar{Y} \\ &= \sum X_i Y_i - \frac{1}{N} \sum X_i \sum Y_i\end{aligned}$$

Also $\sum (X_i - \bar{X})(X_i - \bar{X}) = \sum X_i^2 - \frac{1}{N}(\sum X_i)^2$

Therefore $\hat{\beta}_1 = \frac{S_{XY}}{S_{XX}}$

The OLS solution

- Alternatively:

$$\hat{\beta}_1 = \frac{\sum x_i y_i}{\sum x_i^2} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2}$$

- $(\hat{\beta}_0, \hat{\beta}_1)$ sample estimates of the population parameters (β_0, β_1)
  $\hat{\beta}_0 + \hat{\beta}_1 X \quad \longrightarrow$ estimated line/ sample regression line.
- $\hat{\beta}_1$ cannot be computed if $\sum (X_i - \bar{X})^2 = 0$ i.e. all the X_i's are same. This leads to an important assumption that cannot be relaxed (*unless* $\beta_0 = 0$)

Sample variance – $\widehat{var(X)} = \frac{1}{N-1} \sum (X_i - \bar{X})^2 \neq 0$