Introduction to Big Data and Machine Learning
OLS matrix derivation

Dr. Mihail

August 26, 2019

Ordinary least squares

Matrix form

- Let X be n x k, where each row (n of them) is an observation of k variables. We will assume models have a constant (bias), so first column will be 1's
- Let y be an n x 1 vector of observations on the dependent variable
- Let ϵ be an n x 1 vector of disturbances or errors
- Let β be a k x 1 vector of unknown population parameters that we wish to estimate

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ \vdots \\ Y_n \end{bmatrix}_{nx1} = \begin{bmatrix} 1 & X_{11} & X_{21} & \dots & X_{21} \\ 1 & X_{12} & X_{22} & \dots & X_{k2} \\ \vdots & \vdots & \vdots & \dots & \vdots \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 1 & X_{1n} & X_{2n} & \dots & X_{kn} \end{bmatrix}_{nxk} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \vdots \\ \beta_n \end{bmatrix}_{kx1} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \vdots \\ \epsilon_n \end{bmatrix}_{nx1} \quad (1)$$

Ordinary least squares

Matrix form

$$
\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ \vdots \\ Y_n \end{bmatrix}_{nx1} = \begin{bmatrix} 1 & X_{11} & X_{21} & \dots & X_{21} \\ 1 & X_{12} & X_{22} & \dots & X_{k2} \\ \vdots & \vdots & \vdots & \dots & \vdots \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 1 & X_{1n} & X_{2n} & \dots & X_{kn} \end{bmatrix}_{nxk} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \vdots \\ \beta_n \end{bmatrix}_{kx1} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \vdots \\ \epsilon_n \end{bmatrix}_{nx1}
$$

Or more succintly

$$
y = X\beta + \epsilon \tag{2}
$$

Ordinary least squares

Matrix form

- We wish to estimate $\hat{\beta}$
- $\hat{\beta}$ minimizes the sum of the squared residuals $\sum e_i^2$
- The vector of residuals is given by $e = y - X\hat{\beta}$
- The sum of squared residuals is given by $e'e$[a]

[a] Not to be confused with ee', the covariance of residuals

Sum of squared residuals

$$\begin{bmatrix} e_1 & e_2 & \dots & \dots & e_n \end{bmatrix}_{1xn} \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ \vdots \\ e_n \end{bmatrix} = \begin{bmatrix} e_1 \times e_1 & e_2 \times e_2 & \dots & e_n \times e_n \end{bmatrix} \quad (3)$$

Ordinary least squares

Sum of squares

$$
\begin{aligned}
e'e &= (y - X\hat{\beta})'(y - X\hat{\beta}) \\
&= y'y - \hat{\beta}'y - y'X\hat{\beta} + \hat{\beta}'X'X\hat{\beta} \\
&= y'y - 2\hat{\beta}'X'y + \hat{\beta}'X'X\hat{\beta}
\end{aligned} \tag{4}
$$

We used this identity: $y'X\hat{\beta} = (y'X\hat{\beta})' = \hat{\beta}'X'y$

Ordinary least squares

Matrix differentiation review

$$\frac{\partial a'b}{\partial b} = \frac{\partial b'a}{\partial b} = a \tag{5}$$

where a and b are Kx1 vectors

$$\frac{\partial b'Ab}{\partial b} = 2Ab = 2b'A \tag{6}$$

where A is any symmetric matrix. Note that you can write the derivative as $2Ab$ or $2b'a$

Ordinary least squares

Matrix differentiation review

$$\frac{\partial 2\beta' X' y}{\partial b} = \frac{\partial 2\beta' (X' y)}{\partial b} = 2X' y \tag{7}$$

and

$$\frac{\partial 2\beta' X' X\beta}{\partial b} = \frac{\partial 2\beta' A\beta}{\partial b} = 2A\beta = 2X' X\beta \tag{8}$$

when $X'X$ is a KxK matrix.

Ordinary least squares

Parameter estimation

The $\hat{\beta}$ that minimizes the sum of squared residuals is obtained by computing the derivative of $e'e$ with respect to $\hat{\beta}$

$$\frac{\partial e'e}{\partial \hat{\beta}} = -2X'y + 2X'X\hat{\beta} \tag{9}$$

Setting the derivative equal to 0 and solving for $\hat{\beta}$

$$-2X'y + 2X'X\hat{\beta} = 0 \tag{10}$$

$$(X'X)\hat{\beta} = X'y \tag{11}$$

$X'X$ is always square (k x k) and symmetric.
Both X and y are known from our data

Ordinary least squares

Parameter estimation

$$(X'X)\hat{\beta} = X'y \tag{12}$$

$X'X$ is always square (k x k) and symmetric.
Both X and y are known from our data, so we can multiply both sides by the inverse $(X'X)^{-1}$, yielding:

$$(X'X)^{-1}(X'X)\hat{\beta} = (X'X)^{-1}X'y \tag{13}$$

$$I\hat{\beta} = (X'X)^{-1}X'y \tag{14}$$

or finally:

$$\hat{\beta} = (X'X)^{-1}X'y \tag{15}$$