Lecture 13 Vector Geometry of Linear Models

Objectives

- Simple Linear regression in terms of Linear vector geometry
- Linear model and Fitted Linear model visualization
- Variables in Mean deviation form

Simple Linear Regression can be represented as a vector by the following equation:

$y = \alpha 1_n + \beta x + \varepsilon$ 1

Where $y = [Y_1, Y_2 Y, ... Y_n]'$ $x = [x_1, x_2, x_3 .... x_n]'$ $\varepsilon = [\varepsilon_1, \varepsilon_2, ... \varepsilon_n]'$ $1_n = [1,1,1, ... 1]'$

The aforementioned is the equation for the population with the same assumptions as:

$\varepsilon \sim N_n(0, \sigma_\varepsilon^2 I_n)$ 2

The Fitted Model can be represented as follows:

$y = A1_n + Bx + e$ 3

Where

$e = [E_1, E_2, ... E_n]$ A and B are least square regression coefficient.

$E(y) = A1_n + Bx$ 4

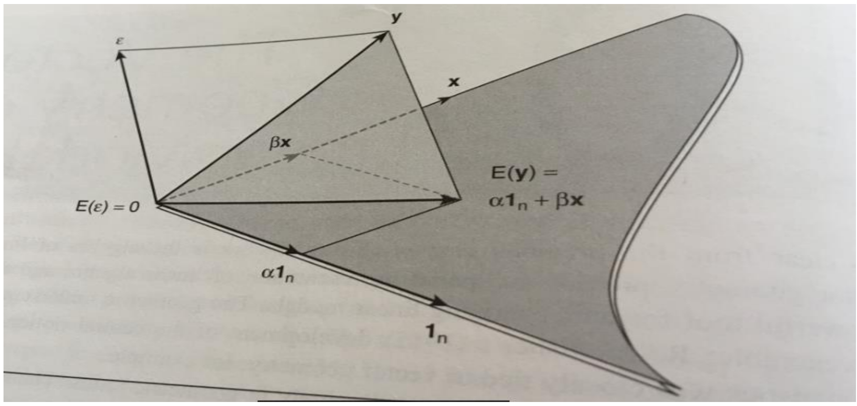
and

$\hat{y} = A1_n + Bx$ 5

For representing this we require an n dimension space. The visualization of n dimensional space cannot be depicted therefore a sub space composed of three dimensional space is used to showcase /conceptualize the geometric properties of of the simple regression model.

The subspace is spanned by x, y, 1_n. y is a random variable that varies from sample to sample therefore here we are only considering one sample.

$E(y) = A1_n + Bx$ is a linear combination of 1_n and x therefore lies in the plane $\{1_n, x\}$.

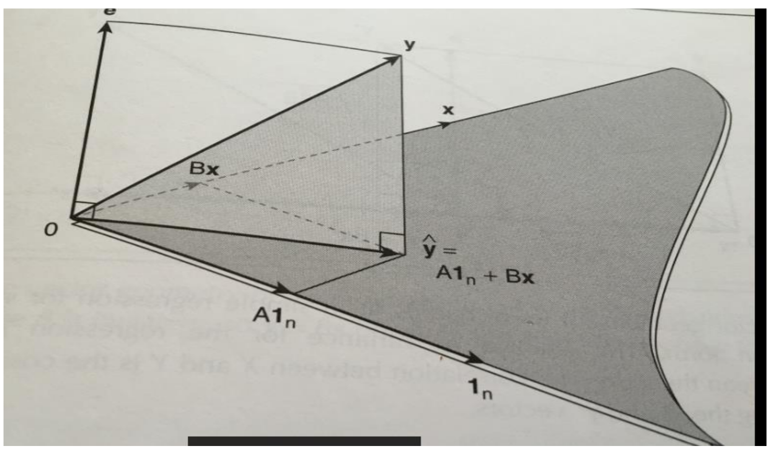


LEAST SQUARE MODEL (Population)

$\varepsilon = y - \alpha 1_n - \beta x$ is non zero in the sample but E(ε) =0 over many samples…………………6

The Simple Linear Regression model is composed of a three dimensional vector space spanned by vectors x, y,1_n The E(y) expected value of y lies in the plane spanned by 1_n and x because the expected error is zero ie E(ε) =0.

The square regression model can be depicted by the following graph:



Linear Least Square Fit

The fitted value $\hat{y}$ is a linear combination of x and 1_n therefore it lies in the plane spanned by 1_n and x ie {1_n, x}. Residual in terms of a vector ie the residual error can be mathematically interpreted as

$e = y - \hat{y}$ has the length (Euclidean norm) in n dimensions $\| e \| = \sqrt{\sum E_i^2}$ …………………7

Remember we are always trying to minimize the Residual sum of square therefore we will endeavor to do this geometrically as well.

Therefore geometrically we try to minimize $\| e \| = \sqrt{\sum E_i^2}$

The error is $e = y - \hat{y}$ is the length between y and $\hat{y}$ .The minimum length is the orthogonal projection of y onto the plane {1_n, x} which is $\hat{y}$. This minimizes $\| e \| = \sqrt{\sum E_i^2}$

Remember the normal equations are :

$$\sum_{i=1}^{n} (A + BX_i - Y_i) = 0$$

$$A\sum_{i=1}^{n} 1 + B\sum_{i=1}^{n} X_i - \sum_{i=1}^{n} Y_i = 0$$

This can be written in vector form by the knowledge that the inner product between two orthogonal vectors =0 as well as that $< u, v > = \sum_{i=1}^{n} u_i v_i$

$$< \vec{1} A + B\vec{X} - \vec{Y}, 1 > = 0$$

Ie <e,1>=0

The second normal equation can be written as

$$\sum_{i=1}^{n} X_i (Y_i - A - BX_i) = 0$$

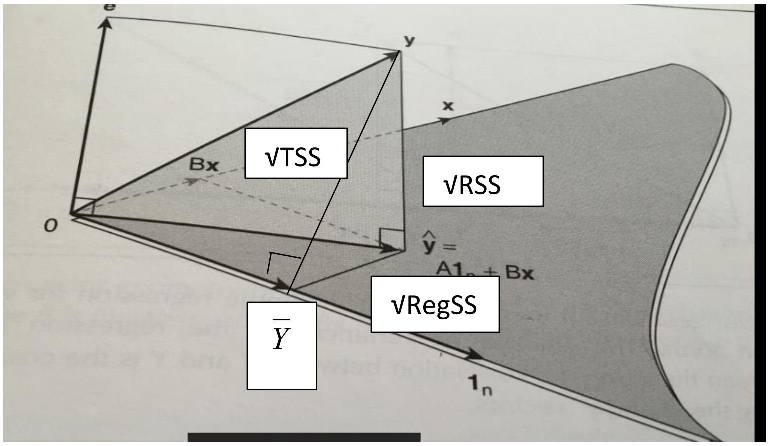
$$\sum_{i=1}^{n} X_i A + \sum_{i=1}^{n} X_i^{\,2} B - \sum_{i=1}^{n} X_i Y_i = 0$$

This can be written in vector form by the knowledge that the inner product between two orthogonal vectors =0 as well as that $< u, v > = \sum_{i=1}^{n} u_i v_i$

$$< \vec{1} A + B\vec{X} - \vec{Y}, X_i > = 0$$

<e,x>=0

Therefore these equations are called normal equations.

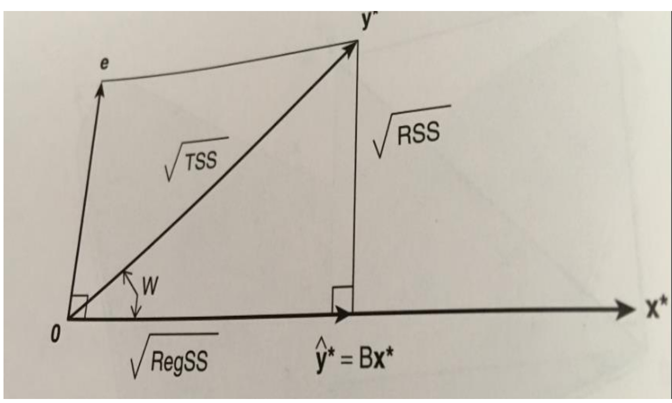


$< \vec{1} A + B\vec{X} - \vec{Y}, 1 >= 0$ when B=0

$< \vec{1} A + \vec{Y}, 1 >= 0$ This projection is mean of Y ie if we fit model just by a line ie a horizontal line for Y mean.

Variables in Mean Deviaton Form

To simplify the graphical representation we can eliminate constant regressor 1_n as well as the intercept coefficient A. This will result the vector space to get transformed into a two dimensional space. We can visually represent ANOVA for regression as well as visually represent multiple regression.



Remember we had eliminated A by setting and manipulating the following equation as follows:

$Y_i = A + Bx_i + E_i$ ……………………………8

As well as $Y = A + B\bar{x}$ ………………………..9

Subtracting 9 from 8

$Y_i - \bar{Y} = B(x_i - \bar{x}) + E_i$ ………………………..10

Setting $y^* \equiv Y_i - \overline{Y}$ and $x^* \equiv x_i - \overline{x}$ in equation 10

$y^* = Bx^* + e$................................11

Remember The fitted value is the deviation of the data from the mean of the data (response variable).

$\hat{y}* \equiv \hat{Y}_i - \overline{Y}$ is a multiple of x (look at the equation 11).

The length of e is minimized by taking an projection of y^* on x^*. This orthogonal projection is $\hat{y}*$

By The formula for orthogonality:

If u and v are vectors then the u* id the projection then

$$u* = \frac{u.v}{\| v \|^2} v$$

$$Bx* = \frac{x*.y*}{\| x \|^2} x*$$

$$B = \frac{x^* y^*}{\| x^* \|^2} = \frac{\sum (x_i - \overline{x})(Y_i - \overline{Y})}{\sum (x_i - \overline{x})^2}$$

$$RSS = \sum E_i^2 = \| e \|^2$$

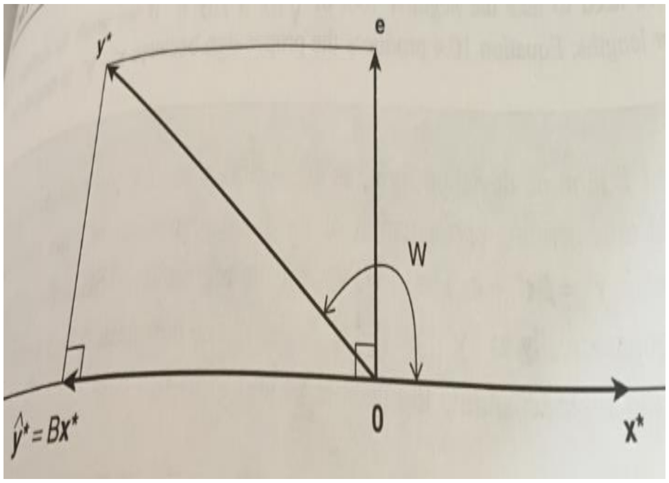
$$TSS = \sum (Y_i - \overline{Y})^2 = \| y^* \|^2$$

And $\operatorname{Re} gSS = \sum (\hat{Y}_i - \overline{Y})^2 = \| \hat{y}* \|^2$

TSS= RegSS + RSS

Correlation Coefficient:

$$r = \sqrt{\frac{\operatorname{Re} gSS}{TSS}} = \frac{\| \hat{y}* \|}{\| y* \|}$$



$$r = \frac{x^* y^*}{\| x^* \| \| y^* \|} = \frac{\sum (x_i - \bar{x})(Y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (Y_i - \bar{Y})^2}}$$