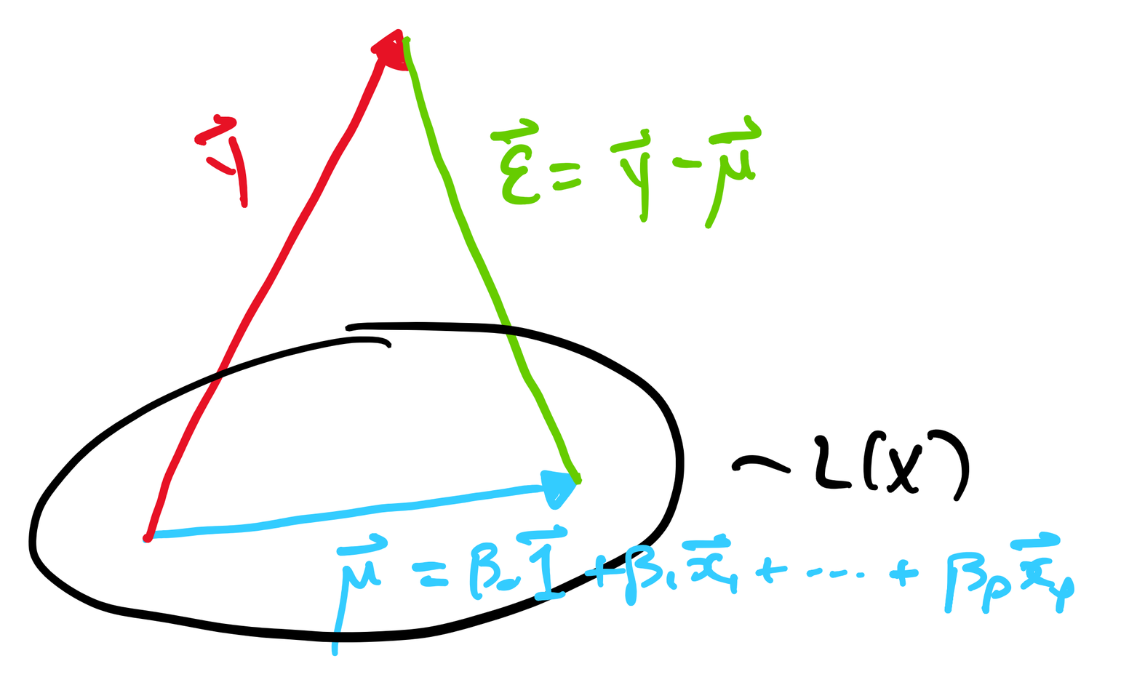
Lecture 8

Wednesday, June 5, 2019 2:36 PM



The goal is to find values of the β's that minimize the magnitude of $\vec{\varepsilon}$.

$$\|\vec{\varepsilon}\|_2 = \sqrt{\vec{\varepsilon}^T \vec{\varepsilon}} = \sqrt{\sum_{i=1}^{n} \varepsilon_i^2} = \sqrt{S(\beta_0, \beta_1, \ldots, \beta_p)}$$

L_2-norm

Because the square root function increases monotonically then the values of the β's that minimize $\|\vec{\varepsilon}\|_2$ are the same ones that minimize $S(\beta_0, \beta_1, \ldots, \beta_p) = \sum_{i=1}^{n} \varepsilon_i^2$. Thus, what we've worked out geometrically, is equivalent to least squares minimization.

The vector in $L(X)$ that is closest to $\vec{y}$ is obtained by making the difference $\vec{y} - \vec{\mu} = \vec{\varepsilon}$ perpendicular (orthogonal) to $L(X)$. Thus we require $\vec{y} - \vec{\mu} = \vec{\varepsilon}$ to be orthogonal to every vector in $L(X)$. In particular, to $\vec{1}, \vec{x}_1, \vec{x}_2, \ldots, \vec{x}_p$. This gives:

$$\vec{1}^T(\vec{y} - \vec{\mu}) = 0$$
$$\vec{x}_1^T(\vec{y} - \vec{\mu}) = 0$$
$$\vec{x}_2^T(\vec{y} - \vec{\mu}) = 0$$
$$\vdots$$
$$\vec{x}_p^T(\vec{y} - \vec{\mu}) = 0$$

As before, this system of equations can written as:

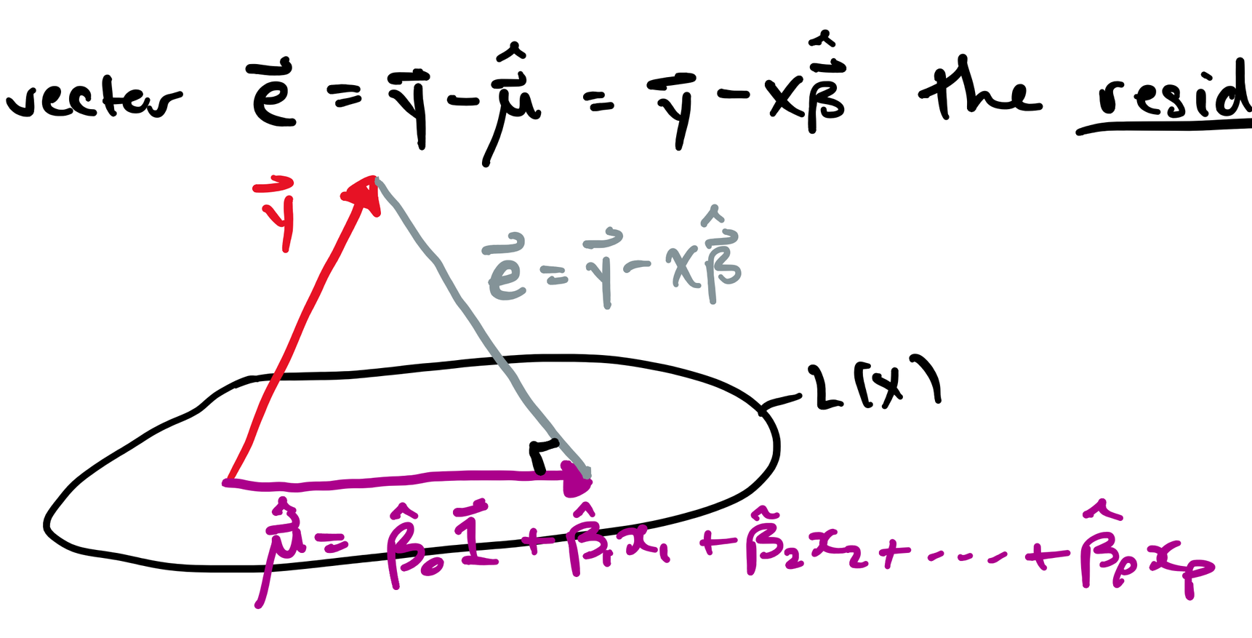
$$X^T(\vec{y} - \vec{\mu}) = \vec{0}$$

As before, solving this equation yields $\hat{\vec{\beta}} = (X^TX)^{-1}X^T\vec{y}$.

__Related Quantities__

We call the vector $\hat{\vec{\mu}} = X\hat{\vec{\beta}}$ the __fitted values__.

We call the vector $\vec{e} = \vec{y} - \hat{\vec{\mu}} = \vec{y} - X\hat{\vec{\beta}}$ the __residuals__.



* The fitted value vector is the __orthogonal projection__ of $\vec{y}$ onto $L(X)$.

The residuals, as in the case of SLR, are used to estimate σ^2, the error variance:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-p-1} = \frac{\vec{e}^T\vec{e}}{n-p-1}$$

$$\hat{\sigma} = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n-p-1}} = \sqrt{\frac{\vec{e}^T\vec{e}}{n-p-1}}$$

($n-p-1$:) "residual degrees of freedom"
= (sample size) − (# of estimated parameters)
$= n - (p+1)$
$= n - p - 1$

__Inference for $\hat{\vec{\beta}} = (X^TX)^{-1}X^T\vec{y}$__

Since $\vec{y} \sim MVN(\vec{\mu} = X\vec{\beta}, \sigma^2 I_{n\times n})$ and because $\hat{\vec{\beta}}$ is a matrix multiple of $\vec{y}$, $\hat{\vec{\beta}}$ also follows a multivariate normal distribution.

- $E[\hat{\vec{\beta}}] = E[(X^TX)^{-1}X^T\vec{y}] = (X^TX)^{-1}X^T E[\vec{y}]$
  $= (X^TX)^{-1}X^TX\vec{\beta}$
  $= \vec{\beta}$

$$\therefore E\begin{bmatrix}\hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_p\end{bmatrix} = \begin{bmatrix}\beta_0 \\ \beta_1 \\ \vdots \\ \beta_p\end{bmatrix}$$

- $Var[\hat{\vec{\beta}}] = Var[(X^TX)^{-1}X^T\vec{y}]$
  $= (X^TX)^{-1}X^T Var[\vec{y}] [(X^TX)^{-1}X^T]^T$
  $= (X^TX)^{-1}X^T(\sigma^2 I) X[(X^TX)^{-1}]^T$
  $= \sigma^2(X^TX)^{-1}X^TX(X^TX)^{-1}$
  $= \sigma^2(X^TX)^{-1}$

$$\therefore \hat{\vec{\beta}} \sim MVN(\vec{\beta}, \sigma^2 \underbrace{(X^TX)^{-1}}_{V})$$

As a consequence, $\hat{\beta}_j \sim N(\beta_j, \sigma^2 v_{jj})$ $j = 0, 1, 2, \ldots, p$

(v_{jj}:) j^{th} diagonal element of V

$$\therefore \frac{\hat{\beta}_j - \beta_j}{\sigma\sqrt{v_{jj}}} \sim N(0,1)$$

$$\Rightarrow \frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}\sqrt{v_{jj}}} \sim t_{(n-p-1)} \qquad \left(\text{since } \frac{(n-p-1)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{(n-p-1)}\right)$$

We use this result to calculate __test statistics__ and __confidence intervals__

$$H_0: \beta_j = 0 \text{ vs. } H_A: \beta_j \neq 0$$
$$\downarrow$$
$$\text{test statistic: } t = \frac{\hat{\beta}_j}{\hat{\sigma}\sqrt{v_{jj}}}$$

($\hat{\sigma}\sqrt{v_{jj}}$:) $SE(\hat{\beta}_j)$

$$\text{p-value} = 2P(t_{(n-p-1)} \geq |t|)$$

A $(1-\alpha)\times 100\%$ CI for β_j is

$$\hat{\beta}_j \pm t_{(n-p-1)}(1-\alpha/2) \times \hat{\sigma}\sqrt{v_{jj}}$$

($\hat{\sigma}\sqrt{v_{jj}}$:) $SE(\hat{\beta}_j)$

where $P\left(W \leq t_{(n-p-1)}(1-\alpha/2)\right) = 1 - \frac{\alpha}{2}$ where $W \sim t_{(n-p-1)}$

__Interpretting the β's__

$$E[y] = \mu = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

Notice that $E[y \mid x_1 = x_2 = x_3 = \cdots = x_p = 0] = \beta_0$. Therefore, β_0 is interpreted as the __expected response__ when all of the explanatory variables are equal to zero. And then, we interpret $\hat{\beta}_0$ as the estimated expected response when all explanatory variables are equal to zero.

Next, notice $E[y \mid x_j = x] = \beta_0 + \beta_1 x_1 + \cdots + \beta_j x + \cdots + \beta_p x_p$

$$E[y \mid x_j = x+1] = \beta_0 + \beta_1 x_1 + \cdots + \beta_j(x+1) + \cdots + \beta_p x_p$$

So $E[y \mid x_j = x+1] - E[y \mid x_j = x] = \beta_j$. Therefore β_j is interpreted as the __expected change__ in the response corresponding to a unit increase in x_j, __when all other explanatory variables stay the same__. And then, we interpret $\hat{\beta}_j$ as the estimate of the expected change in y for a unit increase in x_j, all else held constant.

this is new