Linear Regression

- Let $f(\vec{x};\vec{\theta}) = \underset{\sim}{X}\vec{\theta}$, where $\underset{\sim}{X} \in \mathbb{R}^{N_y \times p}$ + $\vec{\theta} \in \Gamma \subseteq \mathbb{R}^p$.

- Let $\vec{\theta}^0$ be the true, unknown value of parameters.

- Lets assume the following:

$$\left.\begin{aligned} &i)\ E[\varepsilon_i] = 0 \quad , i=1,\dots,N_y \\ &ii)\ Var[\varepsilon_i] = \sigma_0^2 , \quad i=1,\dots,N_y \\ &iii)\ Cov[\varepsilon_i, \varepsilon_j] = 0 , \quad if\ i \neq j \end{aligned}\right\} \overset{iid}{\sim} N(0, \sigma_0^2)$$

Goal: get estimators, $\hat{\vec{\theta}}$ + $\hat{\sigma}^2$, for $\vec{\theta}^0$ + σ_0^2, + then estimates, $\vec{\theta}_{OLS}$ + S^2, with their sampling distributed.

OLS

$$\left(f(\vec{x};\vec{\theta}) = \underset{\sim}{X}\vec{\theta} \right)$$

- Assuming $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$, we seek $\vec{\theta}$ s.t.

$$J(\vec{\theta}) = (\vec{y} - \underset{\sim}{X}\vec{\theta})^T (\vec{y} - \underset{\sim}{X}\vec{\theta})$$

- We want to minimize $J(\vec{\theta})$.

- For vector valued $\vec{\theta} \in \mathbb{R}^p$, we can use the gradient to minimize $J(\vec{\theta})$.

$$\nabla_{\vec{\theta}} J = \nabla_{\vec{\theta}} \left[(\vec{y} - \underset{\sim}{X}\vec{\theta})^T (\vec{y} - \underset{\sim}{X}\vec{\theta}) \right]$$

$$\nabla_{\vec{\theta}} J = \nabla_{\vec{\theta}} \left[(\vec{y} - \underset{\sim}{X}\vec{\theta})^T (\vec{y} - \underset{\sim}{X}\vec{\theta}) \right]$$

$$= -2\underset{\sim}{X}^T \left[\vec{y} - \underset{\sim}{X}\vec{\theta} \right] = -2\left[\underset{\sim}{X}^T \vec{y} - \underset{\sim}{X}^T \underset{\sim}{X} \vec{\theta} \right]$$

- Setting $\nabla_{\vec{\theta}} J = 0$

$$\Rightarrow \boxed{\vec{\theta}_{OLS} = (\underset{\sim}{X}^T \underset{\sim}{X})^{-1} \underset{\sim}{X}^T \vec{y}}$$

- Note: if $\underset{\sim}{X}^T \underset{\sim}{X}$ is ill-conditioned, we can use

$$\underset{\sim}{X}^T \underset{\sim}{X} \approx (\underset{\sim}{X}^T \underset{\sim}{X} + \alpha \underset{\sim}{I})$$

where α regularizes the problem.

Estimator Properties

- Once we get estimates for $\vec{\theta}$, we can look at its sampling distribution. (Note that ^ notation => estimator)

i) $$E[\hat{\theta}] = E\left[(\underset{\sim}{X}^T \underset{\sim}{X})^{-1} \underset{\sim}{X}^T \vec{y} \right] = (\underset{\sim}{X}^T \underset{\sim}{X})^{-1} \underset{\sim}{X}^T E[\vec{y}]$$

$$= (\underset{\sim}{X}^T \underset{\sim}{X})^{-1} \underset{\sim}{X}^T E[\underset{\sim}{X}\vec{\theta} + \vec{\varepsilon}]$$

$$= (\underset{\sim}{X}^T \underset{\sim}{X})^{-1} \underset{\sim}{X}^T \underset{\sim}{X} \vec{\theta}^{\circ} + 0$$

$$= \vec{\theta}^{\circ}$$

ii) $\mathrm{Var}[\hat{\theta}] = E\left[(\hat{\theta} - \vec{\theta}^{\circ})(\hat{\theta} - \vec{\theta}^{\circ})^T \right]$, let $\underset{\sim}{A} = (\underset{\sim}{X}^T \underset{\sim}{X})^{-1} \underset{\sim}{X}^T$

ii) $\mathrm{Var}[\hat{\theta}] = E[(\hat{\theta} - \vec{\theta}^\circ)(\hat{\theta} - \vec{\theta}^\circ)']$, let $A = (X^T X)^{-1} X^T$

(Note: $\hat{\theta} = A y = A(X\theta + \vec{\varepsilon})$)

$$\Rightarrow E\left[(\vec{\theta}^\circ + A\vec{\varepsilon} - \vec{\theta}^\circ)(\vec{\theta}^\circ + A\vec{\varepsilon} - \vec{\theta}^\circ)^T\right]$$

$$= E\left[A \vec{\varepsilon}\, \vec{\varepsilon}^T A^T\right] = A\, E\left[\vec{\varepsilon}\,\vec{\varepsilon}^T\right] A^T$$

$$= A\, \sigma_0^2\, A^T = \boxed{\sigma_0^2 (X^T X)^{-1}}$$

iii) We need an estimator for σ_0^2.

- Let $\hat{R} = \vec{y} - X\hat{\vec{\theta}}$

• Note that $\vec{y} - X\hat{\theta} = \vec{y} - X\left[(X^T X)^{-1} X^T \vec{y}\right]$

$$= (I - H)\vec{y}, \quad H = X(X^T X)^{-1} X^T$$

Note: Since $H = X(X^T X)^{-1} X^T$, $X \in \mathbb{R}^{N_y \times P} \Rightarrow H \in \mathbb{R}^{N_y \times N_y}$

$\Rightarrow H^T = H$ (Sym)

$H^2 = H$ (Idempotent)

$(I - H)^2 = (I - H)$

$(I - H)X = 0$

$$\Rightarrow \hat{R} = (I - H)\vec{y} = (I - H)(X\vec{\theta} + \vec{\varepsilon}) = \underbrace{\cancel{(I - H)X}}_{=0}\vec{\theta} + (I + H)\vec{\varepsilon}$$

$$\Rightarrow \hat{\underset{\sim}{R}} = (\underset{\sim}{I} - \underset{\sim}{H})\vec{y} = (\underset{\sim}{I} - \underset{\sim}{H})(\underset{\sim}{X}\vec{\theta} + \vec{\varepsilon}) = \cancel{(\underset{\sim}{I} - \underset{\sim}{H})\underset{\sim}{X}\vec{\theta}} + (\underset{\sim}{I} + \underset{\sim}{H})\vec{\varepsilon}$$

$$\Rightarrow \hat{\underset{\sim}{R}} = (\underset{\sim}{I} - \underset{\sim}{H})\vec{\varepsilon}$$

- Then $\hat{\underset{\sim}{R}}^{T} \hat{\underset{\sim}{R}} = \vec{\varepsilon}^{T} (\underset{\sim}{I} - \underset{\sim}{H}) \varepsilon$

$$\Rightarrow E\left[\hat{\underset{\sim}{R}}^{T} \underset{\sim}{R}^{T}\right] = E\left[\vec{\varepsilon}^{T} (\underset{\sim}{I} - \underset{\sim}{H})\vec{\varepsilon}\right]$$

$$= \sum_i \sum_j h_{ij} \operatorname{cov}(\varepsilon_i, \varepsilon_j)$$

$$= \sum_i h_{ii} \operatorname{var}(\varepsilon_i)$$

$$= \sigma_0^2 \operatorname{trace}(\underset{\sim}{I} - \underset{\sim}{H})$$

- Note: $\operatorname{trace}(I) = N_y$

$$\operatorname{trace}(H) = \operatorname{trace}\left(\underset{\sim}{X}(\underset{\sim}{X}^T\underset{\sim}{X})^{-1}\underset{\sim}{X}^T\right)$$

- Note: $\operatorname{trace}(\underset{\sim}{A}\underset{\sim}{B}) = \operatorname{trace}(\underset{\sim}{B}\underset{\sim}{A}) = \sum_{i=1}^{N}\sum_{j=1}^{M} a_{ij} b_{ij}$

$$\Rightarrow \operatorname{trace}\left(\underbrace{\underset{\sim}{X}}_{\underset{\sim}{A}}\underbrace{(\underset{\sim}{X}^T\underset{\sim}{X})^{-1}\underset{\sim}{X}^T}_{\underset{\sim}{B}}\right) = \operatorname{trace}\left(\underbrace{(\underset{\sim}{X}^T\underset{\sim}{X})^{-1}\underset{\sim}{X}^T}_{\underset{\sim}{B}}\underbrace{\underset{\sim}{X}}_{\underset{\sim}{A}}\right) = \operatorname{trace}(\underset{\sim}{I}_{N}) = P.$$

$$= \sigma_0^2 \cdot (N_y - P)$$

$$\Rightarrow \hat{\sigma}^2 = \frac{1}{N_y - P} \hat{\underset{\sim}{R}}^{T} \hat{\underset{\sim}{R}}$$

- So in total, given OLS problem with $\vec{y} = \underset{\sim}{X}\vec{\theta} + \vec{\varepsilon}$

$$\boxed{\hat{\vec{\theta}} \sim N\left(\vec{\theta}^{\circ}, \sigma_0^2 \left(\underset{\sim}{X}^T \underset{\sim}{X}\right)^{-1}\right), \quad \sigma_0^2 \approx \frac{1}{N_y - P} \underset{\sim}{R}^T \underset{\sim}{R}}$$

Frequentist Confidence Intervals

- For $N_y \to \infty$, the Law of Large numbers provides an asymptotic framework for $\hat{\vec{\theta}}$ + $\hat{\sigma}^2$
- Since $\hat{\vec{\theta}} \sim N\left(\vec{\theta}^{\circ}, \sigma^2 \left(\underset{\sim}{X}^T \underset{\sim}{X}\right)^{-1}\right)$, then the R.V. T_q is

$$T_q = \frac{\hat{\theta}_q - \theta_q^{\circ}}{\sqrt{\sigma^2 \left(\underset{\sim}{X}^T \underset{\sim}{X}\right)^{-1}_{qq}}}$$

- By law of large numbers, $T_q \sim$ t-distribution with $N_y - P$ degrees of freedom

- Then, $\hat{\theta}_q$ has a $1-\alpha$ confidence interval given by

$$P\left(\rho_-(\hat{\theta}_q) < \hat{\theta}_q < \rho_+(\hat{\theta}_q)\right) = 1 - \alpha$$

$$+ \quad \rho_\pm = \hat{\theta}_q \pm t_{N_y - P}^{1-\alpha/2} \cdot \sqrt{\sigma^2 \left(\underset{\sim}{X}^T \underset{\sim}{X}\right)^{-1}}$$