FYS-STK4155 sept 2

OLS = ORDINARY LEAST SQUARES

$$y = f(x) + \varepsilon$$

$$\varepsilon \sim N(0, \sigma_\varepsilon^2)$$

$$f(x) \simeq X\beta$$

$$y \in \mathbb{R}^n \qquad X \in \mathbb{R}^{n \times p}$$

$$\beta \in \mathbb{R}^p$$

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

$$MSE = C(\beta) = \frac{1}{n} \sum_{i=0}^{n-1} (y_i - \tilde{y}_i)^2$$

$$\tilde{y}_i = \sum_{j=0}^{p-1} x_{ij} \beta_j$$

$$= X_{i*} \beta$$

$$\frac{\partial C(\beta)}{\partial \beta} = 0 = X^T (y - X\beta)$$

$$\frac{\partial^2 C(\beta)}{\partial \beta' \partial \beta} = \frac{2}{n} X^T X$$

Hessian $H = X^T X$

Back to statistics

$$\mathbb{E}[y] = ?$$

$$\mathbb{E}[y] = \frac{1}{n} \sum_{i=0}^{n-1} y_i$$

$$y_i = \underbrace{\sum_{j=0}^{p-1} x_{ij} \beta_j}_{\underbrace{X_{i*} \beta}_{\text{not stochastic}}} + \varepsilon_i$$

$$\mathbb{E}[y_i] = \mathbb{E}[X_{i*} \beta] + \mathbb{E}[\varepsilon_i]$$

$$= X_{i*} \beta + 0$$

$$\mathbb{E}[y] = X \beta$$

$$\mathrm{var}[y_i] = \mathbb{E}\left[(y_i - \mathbb{E}[y_i])^2\right]$$

$$= \mathbb{E}[y_i^2] - \underbrace{(\mathbb{E}[y_i])^2}_{(x_{i*}\beta)^2}$$

$$y_i = x_{i*}\beta + \varepsilon_i$$

$$\mathbb{E}\left[(x_{i*}\beta)^2 + 2\varepsilon_i x_{i*}\beta + \varepsilon_i^2\right]$$

$$= \cancel{(x_{i*}\beta)^2} + 2x_{i*}\beta\,\mathbb{E}[\varepsilon_i]^{=0} +$$

$$\underbrace{\mathbb{E}[\varepsilon_i^2]}_{=\sigma_\varepsilon^2} - \cancel{(x_{i*}\beta)^2}$$

$$\mathrm{var}(y_i) = \sigma_\varepsilon^2$$

$$\Rightarrow \quad p(y_i \mid X\beta) = P_i$$

$$\sim N(x_{i*}\beta, \sigma_\varepsilon^2)$$

$$p(y \mid X\beta) = \prod_{i=0}^{n-1} p_i$$

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

$$= (X^T X)^{-1} X^T \underbrace{\mathbb{E}[y]}_{X\beta}$$

$$= \beta,$$ unbiased estimator/parameter

Exercise week 36

$$\text{var}[\beta] = \sigma_\varepsilon^2 (X^T X)^{-1}$$

standard deviation

$$= \sqrt{\text{var}[\beta]} = \Delta\hat{\beta}$$

$$\hat{\beta} \pm \Delta\hat{\beta}$$

Meet the covariance

Define two vectors

x and y

$x, y \in \mathbb{R}$

$$\underset{\text{(sample)}}{\operatorname{cov}[x,y]} = \frac{1}{n} \sum_{i=0}^{n-1} \underbrace{(x_i - \bar{\mu}_x)}_{\tilde{x}_i} \underbrace{(y_i - \bar{\mu}_y)}_{\tilde{y}_i}$$

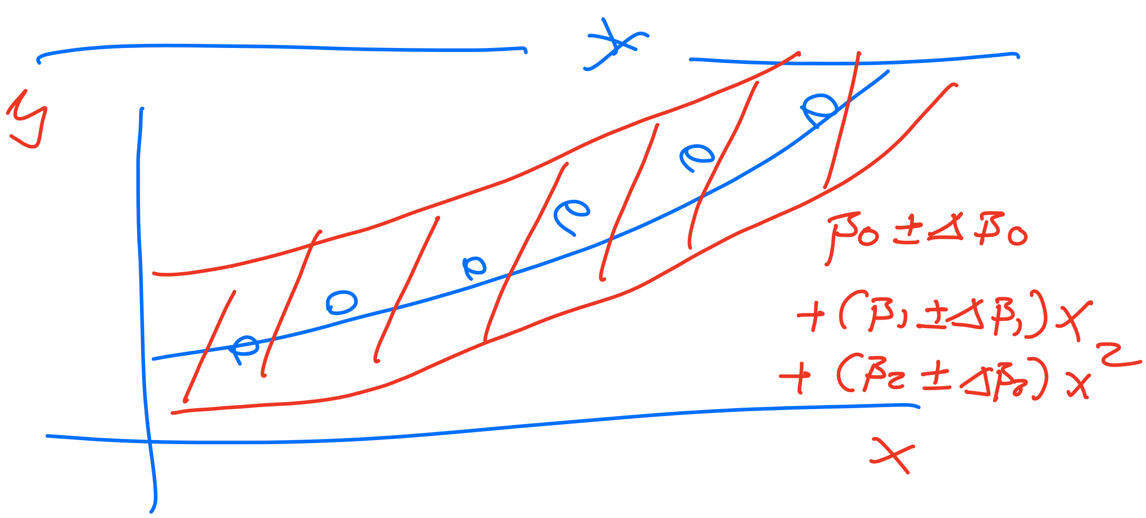
$$= \frac{1}{n} \sum_{i=0}^{n-1} \tilde{x}_i \tilde{y}_i$$

$$\operatorname{var}[x] = \frac{1}{n} \sum_{i=0}^{n-1} \tilde{x}_i^2$$

Define covariance matrix

$$C[x,y] = \begin{bmatrix} \operatorname{cov}[x,x] & \operatorname{cov}[x,y] \\ \operatorname{cov}[y,x] & \operatorname{cov}[y,y] \end{bmatrix}$$

$$= \begin{bmatrix} \operatorname{var}[x] & \operatorname{cov}[x,y] \\ \operatorname{cov}[x,y] & \operatorname{var}[y] \end{bmatrix}$$



Design matrix

$$X = \begin{bmatrix} x_{00} & x_{01} & \cdots & x_{0p-1} \\ x_{10} & & & \\ x_{20} & & & \\ \vdots & & & \\ x_{n-1\,0} & - & - & x_{n-1\,p-1} \end{bmatrix}$$

$$X = \begin{bmatrix} \mathbb{X}_0 & \mathbb{X}_1 & \cdots & \mathbb{X}_{p-1} \end{bmatrix}$$

$$\mathbb{X}_0 = \begin{bmatrix} x_{00} \\ x_{10} \\ \vdots \\ x_{n-1\,0} \end{bmatrix}$$

$$C[X] = \begin{bmatrix} var[\mathbb{X}_0] & cov[\mathbb{X}_0, \mathbb{X}_1] & \cdots & cov[\mathbb{X}_0, \mathbb{X}_{p-1}] \\ & \ddots & & \\ & & \ddots & \\ cov[\mathbb{X}_0\, \mathbb{X}_{p-1}] & \cdots & & var[\mathbb{X}_{p-1}] \end{bmatrix}$$

$$corr[X, Y] = \frac{cov[X, Y]}{\sqrt{var[X]\, var[Y]}}$$

Correlation matrix

$$K[X] = \begin{bmatrix} 1 & \mathrm{corr}[x_0 x_1] & \cdots & \mathrm{corr}[x_0, x_{p-1}] \\ \vdots & 1 & & \\ \vdots & & 1 & \\ \vdots & & & \ddots \\ \mathrm{corr}[x_0, x_{p-1}] & \cdots & \cdots & 1 \end{bmatrix}$$

Ridge regression

$X^T X$ can be non-invertible

Add a small number λ to the diagonal elements.

Example

$$A = \begin{bmatrix} 1 & -1 \\ 1 & -1 \end{bmatrix} \quad \det(A) = 0$$

$$\begin{bmatrix} 1+\lambda & -1 \\ 1 & -1+\lambda \end{bmatrix} \quad \det(A) \neq 0$$

$$C^{OLS}(\beta) = \frac{1}{n}\sum_{i=0}^{n-1}(y_i - \tilde{y}_i)^2$$

$$C^{Ridge}(\beta) = \frac{1}{n}\sum_{i=0}^{n-1}(y_i - \tilde{y}_i)^2 + \lambda\sum_{j=0}^{p-1}\beta_j^2$$

hyperparameter

regularization term.

$$0 \leq \lambda \leq \mathrm{Max}$$

l2 norm

$$\|v\|_2 = \sqrt{\sum_{i=0}^{n-1} v_i^2}$$

$$C^{Ridge}(\beta) = \underbrace{\frac{1}{n}\|y - \tilde{y}\|_2^2}_{MSE} + \lambda\|\beta\|_2^2$$

$$\mathbb{E}\left[(y-\tilde{y})^2\right]$$

l_1-norm $\sum_{j=0}^{p-1} |\beta_j| = \|\beta\|_1$

$$C^{LASSO}(\beta) = \frac{1}{n}\|(y-\tilde{y})\|_2^2 + \lambda\|\beta\|_1$$

$$\frac{\partial C^{Ridge}}{\partial \beta} = -\frac{2}{\cancel{n}} X^T(y - X\beta) + 2\lambda\beta = 0$$

$$\Rightarrow X^T X\beta + \lambda\beta = X^T y \Rightarrow$$

$$\hat{\beta} = (X^T X + \mathbb{I}\lambda)^{-1} X^T y$$

key message:
λ can be used to "shrink away" less important features. Think of λ as a way to reduce the

dimensionality.
The MSE with Ridge and Lasso regression can be tuned down by appropriate λ – values.