Exercise 1 Given data: $\vec{y} = f(x) + \vec{\varepsilon}$ with $\varepsilon \sim \mathcal{N}(0, \sigma^2)$

Approximate or model $f(x)$ as $\tilde{y}(x) = X\vec{\beta}$ then $\vec{y} = X\vec{\beta} + \vec{\varepsilon}$

In components: $y_i = \sum_{j=0}^{P-1} X_{ij}\beta_j + \varepsilon_i = X_{i*}\vec{\beta} + \varepsilon_i$

Therefore $\mathbb{E}[y_i] = \underbrace{\mathbb{E}\left[\sum_{j=0}^{P-1} X_{ij}\beta_j\right]}_{\text{NON-STOCHASTIC}} + \underbrace{\mathbb{E}[\varepsilon_i]}_{=0} = \sum_{j=0}^{P-1} X_{ij}\beta_j = X_{i*}\vec{\beta}$ $\sim N(0, \sigma^2)$

The variance is

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

$$= \mathbb{E}[y_i^2] + \mathbb{E}\left[\mathbb{E}[y_i]^2\right] - 2\,\mathbb{E}\left[y_i\mathbb{E}[y_i]\right]$$

$$= \mathbb{E}[y_i^2] + \mathbb{E}[y_i]^2 - 2\,\mathbb{E}[y_i]\mathbb{E}[y_i] = \mathbb{E}[y_i^2] - \mathbb{E}[y_i]^2$$

$$= \mathbb{E}\left[(X_{i*}\vec{\beta} + \varepsilon_i)^2\right] - \mathbb{E}\left[X_{i*}\vec{\beta} + \varepsilon_i\right]^2$$

$$= \mathbb{E}\left[(X_{i*}\vec{\beta})^2 + 2X_{i*}\vec{\beta}\varepsilon_i + \varepsilon_i^2\right] - \mathbb{E}[y_i]^2$$

$$= (X_{i*}\vec{\beta})^2 + 2X_{i*}\vec{\beta}\,\underbrace{\mathbb{E}[\varepsilon_i]}_{=0} + \mathbb{E}[\varepsilon_i^2] - (X_{i*}\vec{\beta})^2$$

$$= \mathbb{E}[\varepsilon_i^2] \qquad \mathrm{Var}\{\varepsilon_i\} = \mathbb{E}[\varepsilon_i^2] - \cancelto{0}{\mathbb{E}[\varepsilon_i]}^2$$

$$= \sigma^2 \qquad \sigma^2 = \mathbb{E}[\varepsilon_i^2]$$

In OLS the optimal betas are: $\hat{\beta} = [X^T X]^{-1} X^T \vec{y}$

with X non-stochastic.

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

$$= [X^T X]^{-1} X^T X\vec{\beta}$$

$$= \mathbb{1}\vec{\beta}$$

$$= \vec{\beta}$$

The variance is $\mathrm{Var}\{\beta_i;\beta_j\} = \mathbb{E}\left[(\beta_i - \mathbb{E}[\beta_i])(\beta_j - \mathbb{E}[\beta_j])\right]$

In matrix form $\mathrm{Var}\{\vec{\beta}\} = \mathbb{E}\left\{(\vec{\beta} - \mathbb{E}[\vec{\beta}])(\vec{\beta} - \mathbb{E}[\vec{\beta}])^T\right\}$

with $\vec{\beta}_{OLS} = [X^TX]^{-1}X^T\vec{y}$ and $\mathbb{E}[\vec{\beta}] = \vec{\beta}$

$$\Rightarrow \mathrm{Var}\{\vec{\beta}\} = \mathbb{E}\left\{\left[(X^TX)^{-1}X^T\vec{y} - \vec{\beta}\right]\left[(X^TX)^{-1}X^T\vec{y} - \vec{\beta}\right]^T\right\}$$

$$= \mathbb{E}\left\{\left[(X^TX)^{-1}X^T\vec{y} - \vec{\beta}\right]\left[\vec{y}^TX(X^TX)^{-1} - \vec{\beta}^T\right]\right\}$$ where $[(X^TX)^{-1}]^T = [(X^TX)^T]^{-1} = (X^TX)^{-1}$

$$= \mathbb{E}\left\{(X^TX)^{-1}X^T\vec{y}\vec{y}^TX(X^TX)^{-1} - (X^TX)^{-1}X^T\vec{y}\vec{\beta}^T - \vec{\beta}\vec{y}^TX(X^TX)^{-1} + \vec{\beta}\vec{\beta}^T\right\}$$

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

we use: $\vec{y} = X\vec{\beta} + \vec{\varepsilon} \longrightarrow \mathbb{E}[\vec{y}] = X\vec{\beta}$

$\vec{y}^T = \vec{\beta}^TX^T + \vec{\varepsilon}^T \longrightarrow \mathbb{E}[\vec{y}^T] = \vec{\beta}^TX^T$

And $\vec{y}\vec{y}^T = [X\vec{\beta} + \vec{\varepsilon}][\vec{\beta}^TX^T + \varepsilon^T] = X\vec{\beta}\vec{\beta}^TX^T + X\vec{\beta}\epsilon^T + \vec{\varepsilon}\vec{\beta}^TX^T + \vec{\varepsilon}\vec{\varepsilon}^T$

$$\mathbb{E}[\vec{y}\vec{y}^T] = X\vec{\beta}\vec{\beta}^TX^T + X\vec{\beta}\underbrace{\mathbb{E}[\varepsilon^T]}_{=0} + \underbrace{\mathbb{E}[\vec{\varepsilon}]}_{=0}\vec{\beta}^TX^T + \underbrace{\mathbb{E}[\vec{\varepsilon}\vec{\varepsilon}^T]}_{=\sigma^2\mathbb{1}}$$

$$= X\vec{\beta}\vec{\beta}^TX^T + \sigma^2\mathbb{1}$$

Replacing

since $\mathbb{E}[\varepsilon_i\varepsilon_j] = \sigma^2\delta_{ij}$

$$\mathrm{Var}\{\vec{\beta}\} = (X^TX)^{-1}X^T(X\vec{\beta}\vec{\beta}^TX^T + \sigma^2\mathbb{1})X(X^TX)^{-1} - (X^TX)^{-1}X^TX\vec{\beta}\vec{\beta}^T$$
$$- \vec{\beta}\vec{\beta}^TX^TX(X^TX)^{-1} + \vec{\beta}\vec{\beta}^T$$
$$= (X^TX)^{-1}X^TX\vec{\beta}\vec{\beta}^TX^TX(X^TX)^{-1} + \sigma^2(X^TX)^{-1}X^TX(X^TX)^{-1}$$
$$- \vec{\beta}\vec{\beta}^T - \cancel{\vec{\beta}\vec{\beta}^T} + \cancel{\vec{\beta}\vec{\beta}^T}$$
$$= \cancel{\vec{\beta}\vec{\beta}^T} + \sigma^2(X^TX)^{-1} - \cancel{\vec{\beta}\vec{\beta}^T}$$

$$\boxed{\mathrm{Var}\{\vec{\beta}\} = \sigma^2[X^TX]^{-1}}$$

$$\mathrm{Var}\{\beta_K\} = \sigma^2[X^TX]^{-1}_{KK}$$

Exercise 2

In Ridge regression, optimal betas are: $\hat{\beta} = [X^TX + \lambda \mathbb{1}]^{-1} X^T \vec{y}$

Then $E[\hat{\beta}] = E\{[X^TX+\lambda\mathbb{1}]^{-1}X^T\vec{y}\} = [X^TX+\lambda\mathbb{1}]^{-1}X^T \mathbb{E}[\vec{y}]$

$E[\hat{\beta}_{Ridge}] = [X^TX+\lambda\mathbb{1}]^{-1}X^TX\vec{\beta} \neq E[\hat{\beta}_{OLS}]$ for $\lambda > 0$.

The variance is

$$Var\{\hat{\beta}_{Ridge}\} = \mathbb{E}\{(\hat{\beta}_{Rid} - \mathbb{E}(\hat{\beta}_{rid}))(\beta_{rid} - \mathbb{E}(\hat{\beta}_{rid}))^T\}$$

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

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

$$= \mathbb{E}\{(X^TX+\lambda\mathbb{1})^{-1}X^T\vec{y}\vec{y}^TX(X^TX+\lambda\mathbb{1})^{-1} - (X^TX+\lambda\mathbb{1})^{-1}X^T\vec{y}\vec{\beta}^TX^TX(X^TX+\lambda\mathbb{1})^{-1}$$
$$- (X^TX+\lambda\mathbb{1})^{-1}X^TX\vec{\beta}\vec{y}^TX(X^TX+\lambda\mathbb{1})^{-1} + (X^TX+\lambda\mathbb{1})^{-1}X^TX\vec{\beta}\vec{\beta}^TX^TX(X^TX+\lambda\mathbb{1})^{-1}\}$$

We use previous calculations: $\mathbb{E}[\vec{y}\vec{y}^T] = X\vec{\beta}\vec{\beta}^TX^T + \sigma^2\mathbb{1}$, $\mathbb{E}[\vec{y}] = X\vec{\beta}$, $\mathbb{E}[\vec{y}^T] = \vec{\beta}^TX^T$

$$= (X^TX+\lambda\mathbb{1})^{-1}X^T(X\vec{\beta}\vec{\beta}^TX^T + \sigma^2\mathbb{1})X(X^TX+\lambda\mathbb{1})^{-1} - (X^TX+\lambda\mathbb{1})^{-1}X^TX\vec{\beta}\vec{\beta}^TX^TX(X^TX+\lambda\mathbb{1})^{-1}$$
$$- (X^TX+\lambda\mathbb{1})^{-1}X^TX\vec{\beta}\vec{\beta}^TX^TX(X^TX+\lambda\mathbb{1})^{-1} + (X^TX+\lambda\mathbb{1})^{-1}X^TX\vec{\beta}\vec{\beta}^TX^TX(X^TX+\lambda\mathbb{1})^{-1}$$

$$= (X^TX+\lambda\mathbb{1})^{-1}X^TX\vec{\beta}\vec{\beta}^TX^TX(X^TX+\lambda\mathbb{1})^{-1} + \sigma^2(X^TX+\lambda\mathbb{1})^{-1}X^TX(X^TX+\lambda\mathbb{1})^{-1}$$
$$- (X^TX+\lambda\mathbb{1})^{-1}X^TX\vec{\beta}\vec{\beta}^TX^TX(X^TX+\lambda\mathbb{1})^{-1}$$

$$= \sigma^2(X^TX+\lambda\mathbb{1})^{-1}X^TX(X^TX+\lambda\mathbb{1})^{-1}$$

Notice $[(X^TX+\lambda\mathbb{1})^{-1}]^T = [(X^TX+\lambda\mathbb{1})^T]^{-1} = [(X^TX)^T + (\lambda\mathbb{1})^T]^{-1} = (X^TX+\lambda\mathbb{1})^{-1}$

Then we can re-write:

$$Var\{\hat{\beta}_{Ridge}\} = \sigma^2(X^TX+\lambda\mathbb{1})^{-1}X^TX[(X^TX+\lambda\mathbb{1})^{-1}]^T$$