Today

1) 3 exercises
2) Gauss-Markov theorem
   orthogonalization
3) Model selection
4) Shrinkage methods (ch 3.4)

Ex 2.7

N pairs (x_i, y_i) drawn iid from

$$x_i \sim h(x)$$
$$y_i = f(x_i) + \varepsilon_i$$
$$\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$$

Estimator for f linear in the y_i:

$$\hat{f}(x_0) = \sum_{i=1}^{N} \ell_i(x_0; \mathcal{X})\, y_i$$

where weights $\ell_i(x_0; \mathcal{X}) \perp\!\!\!\perp y_i$, but depend on $\mathcal{X}$, the entire training sequence of x_i

a) $$\hat{f}_{LR}(x_0) = X_0^T \hat{\beta} = x_0^T \underbrace{(X^T X)^{-1} X^T y_i}_{\beta}$$

$$\Rightarrow \ell_i(x_0, \mathcal{X}) = \left[x_0^T (X^T X)^{-1} X^T\right]_{[i]}$$

$$\hat{f}_{KNN}(x_0) = \text{Ave}\left\{ y_i \mid x_i \in N_k(x_0) \right\}$$

where $N_k(x_0)$ is the set of the k nearest neighbors

$$\Rightarrow \ell_i(x_0, \mathcal{X}) = \begin{cases} 1/k & \text{if } x_i \in N_k(x_0) \\ 0 & \text{otherwise} \end{cases}$$

b) Decompose the conditional mean-squared error

$$E_{Y|X}\left[\left(f(x_0)-\hat f(x_0)\right)^2\right]$$

stochastic fixed

$$E_{Y|X}\left[\left(\underbrace{f(x_0)-E_{Y|X}[\hat f(x_0)]}+\underbrace{E_{Y|X}[\hat f(x_0)]-\hat f(x_0)}\right)^2\right]$$

$E_{Y|X}[f(x_0)-E_{Y|X}[\hat f(x_0)]]^2$

$$=E_{Y|X}\left[\left(f(x_0)-E_{Y|X}[\hat f(x_0)]\right)^2+\left(E_{Y|X}[\hat f(x_0)]-\hat f(x_0)\right)^2\right.$$
$$\left.+2\left(f(x_0)-E_{Y|X}[\hat f(x_0)]\right)\left(E_{Y|X}[\hat f(x_0)]-\hat f(x_0)\right)\right]$$

$$=\underbrace{\left(f(x_0)-E_{Y|X}[\hat f(x_0)]\right)^2}_{\text{bias}^2}+\underbrace{Var_{Y|X}\left(\hat f(x_0)\right)}_{\text{variance}}+$$
$$+2\left(f(x_0)-E_{Y|X}[\hat f(x_0)]\right)E_{Y|X}\Big[\underbrace{E_{Y|X}[\hat f(x_0)]-\hat f(x_0)}_{=0}\Big]$$

$$=\text{bias}^2+\text{variance}$$

c) unconditional

$$E_{Y,X}\left[\left(f(x_0)-\hat f(x_0)\right)^2\right]$$

$$E_X\left[E_{Y|X}\left[\left(f(x_0)-\hat f(x_0)\right)^2\right]\right]$$

d) conditional:

$$\left(f(x_0)-E_{Y|X}[\hat f(x_0)]\right)^2+Var_{Y|X}\left(\hat f(x_0)\right)$$

$$\left(f(x_0)-E_{Y|X}\left[\sum l_i(x_0;\mathcal{X})\,y_i\right]\right)^2+Var_{Y|X}\left[\sum_{i=1}^{\tilde n} l_i(x_0;\mathcal{X})\,y_i\right]$$

$$\left(f(x_0)-\sum_{i=1}^{\tilde n} l_i(x_0;\mathcal{X})\,f(x_0)\right)^2+\sum_{i=1}^{\tilde n} l_i(x_0;\mathcal{X})^2\,\sigma^2$$

unconditional

$$E_X\left[\left(f(x_0)-E_{Y|X}\left[\sum l_i(x_0;\mathcal{X})y_i\right]\right)^2\right]+E_X\left[Var_{Y|X}\left[\sum_{i=1}^{\tilde n} l_i(x_0;\mathcal{X})y_i\right]\right]$$

Ex 3.1

$$F = \frac{(RSS_0 - RSS_1)/(p_1 - p_0)}{RSS_1/(N - p_1 - 1)}$$

when we are testing only one parameter $F \approx z^2$

$\rightarrow p_1 = p, \quad p_0 = p - 1$

$$F = \frac{(RSS_0 - RSS_1)}{RSS_1/(N - p - 1)} \sim F_{1, N-p-1}$$

$$z = \frac{\hat{\beta}_j}{\hat{\sigma}[X^T X]_{(jj)}} \sim t_{N-p-1}$$

$$(t_{N-p-1})^2 \overset{d}{=} F_{1, N-p-1} \Rightarrow z^2 \approx F$$

Gauss - Markov theorem

$\hat{\beta}_{LS}$ is BLUE

Best $\leftarrow$ smallest variance

Linear $\leftarrow$ $\hat{\vartheta} = a^T \hat{\beta} \leftrightarrow \vartheta = a^T \beta$

Unbiased $\leftarrow$ $E[\hat{\vartheta}] = \vartheta$

Estimator

consider X fixed

$$E_{Y|X}[a^T \hat{\beta}] = E_{Y|X}[a^T (X^T X)^{-1} X^T y] \overset{!}{=} a^T \underbrace{(X^T X)^{-1} X^T X}_{I} \beta \overset{!}{=} a^T \beta$$

Th: $\mathrm{Var}(a^T \hat{\beta}) \leq \mathrm{Var}(c^T \hat{\beta})$ $\rightarrow$ ex 3.3 (a)

$$MSE(\tilde{\vartheta}) = E[(\tilde{\vartheta} - \vartheta)^2] \overset{!}{=} E[(\tilde{\vartheta} - E[\tilde{\vartheta}] + E[\tilde{\vartheta}] - \vartheta)^2] \overset{!}{=} \mathrm{Var}(\tilde{\vartheta}) + (\underbrace{E[\tilde{\vartheta}] - \vartheta}_{=0})^2$$

smallest with LS

- $MSE(\hat{\theta})$ is strongly related with the prediction error

Let consider $Y_0 = f(x_0) + \varepsilon \qquad \varepsilon \sim N(0,\sigma^2)$

$\varepsilon \perp x$

$$E[(Y_0 - \hat{f}(x_0))^2] = E[Y_0^2 + \hat{f}(x_0)^2 - 2Y_0\hat{f}(x_0)]$$

$$= E[Y_0^2] + E[\hat{f}(x_0)^2] - 2E[Y_0\hat{f}(x_0)]$$

$$= Var(Y_0) + \cancel{E[Y_0]}^{2} + Var(\hat{f}(x_0)) + E[\hat{f}(x_0)]^2 - 2f(x_0)E[\hat{f}(x_0)]$$

$f(x_0)$

$$= \sigma^2 + Var(\hat{f}(x_0)) + \underbrace{E[\hat{f}(x_0)]^2 - 2f(x_0)E[\hat{f}(x_0)] + f(x_0)^2}_{(E[\hat{f}(x_0)] - Y_0)^2}$$

$$= \sigma^2 + \underbrace{Var(\hat{f}(x_0)) + bias^2}_{MSE}$$

$$E[Y_0] = E[f(x_0) + \varepsilon] = E[f(x_0)] + E[\varepsilon] = f(x_0) + 0$$

$$E[Y_0^2] = E[(f(x_0) + \varepsilon)^2] = Var[Y_0] + E[Y_0]^2 = \sigma^2 + f(x_0)^2$$

Suppose $Y = X\beta + \varepsilon$

univariable

if $Y = X_1 \beta_1 + \varepsilon$

$$\hat{\beta} = \frac{\sum x_i y_i}{\sum x_i^2} = \frac{\langle x, y \rangle}{\langle x, x \rangle}$$

multivariable

if $Y = X_1 \beta_1 + X_2 \beta_2 + \varepsilon$

only if X is orthogonal, $\hat{\beta}_j = \frac{\langle x_j, y \rangle}{\langle x_j, x_j \rangle}$

alternative formula for variance of $\hat{\beta}_j$

$$Var(\hat{\beta}_j) = \frac{\sigma^2}{\langle z_p, z_p \rangle} = \frac{\sigma^2}{\|z_p\|^2}$$

→ variables may share the same information useful for explaining/predicting y

→ the estimates may be unstable due to high variance

↓

model selection

$$z_{score} = \frac{\hat{\beta}}{sd(\hat{\beta})}$$

→ smaller

→ larger

Model Selection

→ prediction accuracy
→ interpretability
→ portability

• best subset technique $\qquad X = (x_1, x_2, x_3$

models : 0 variables $\quad y = \beta_0 + \varepsilon$

1 variable $\quad y = \beta_0 + \beta_1 x_1 + \varepsilon$

$y = \beta_0 + \beta_2 x_2 + \varepsilon$

$y = \beta_0 + \beta_3 x_3 + \varepsilon$

2 variables $\quad y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$

$y = \beta_0 + \beta_1 x_1 + \beta_3 x_3 + \varepsilon$

$\binom{3}{2} = \frac{3 \cdot \cancel{2}}{\cancel{2}} = 3$ → $y = \beta_0 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$

→ $\binom{8}{2} = \frac{8 \cdot 7}{2} = 28$ picture

• Stepwise techniques

– forward selection

start : $y = \beta_0 + \varepsilon \qquad x_1 \quad x_2 \quad x_3$

1^{st} step : $y = \beta_0 + \beta_2 x_2 + \varepsilon \qquad x_1 \quad x_3$

2^{nd} step : $y = \beta_0 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$

$\hat{\beta}_3$ together with $\hat{\beta}_0$ and $\hat{\beta}_2$

- backward elimination

start : full model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$$

p>>n case backward elimination is not possible

1^{st} step $\quad y = \beta_0 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$

⋮

- stepwise selection

stepback "

$$RSS(\theta) + J(\theta)$$
$$+ 2|p|$$

---

stagewise regression

$$y = \beta_0 + \beta_1 x_1 + \varepsilon$$
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

stepwise: the estimate of β take into account all x

stagewise: introduce a new β, its estimate in only based on x_j.