STATS 315A Individual Problems

2.

- We have

$$z_i = a^T x_i \text{ , so } E[z_i] = E[a^T x_i]$$

$$= a^T E[x_i] = a^T \vec{0} = \boxed{0}$$

Similarly, $Var[z_i] = Var[a^T x_i]$

$$= a^T Var[x_i] a = a^T I_p a.$$

$$= a^T a = \frac{x_0^T x_0}{||x_0||^2} = \boxed{1}$$

As normality is preserved under linear transformations,

$$z_i \sim N(E[z_i], Var[z_i])$$

$$= N(0,1)$$

- $E[z_i^2] = Var[z_i] + (E[z_i])^2$

$$= \boxed{1}$$

$$\therefore E[Z^T Z] = E\left[\sum_{i=1}^{p} z_i^2\right] = \sum_{i=1}^{p} E[z_i]^2$$

target point

$$= \boxed{p}$$

3 a.)

$$E[\hat{\beta} \mid X] = E[(X^TX)^{-1}X^Ty \mid X]$$

$$= E[(X^TX)^{-1}X^T(X\beta+\varepsilon) \mid X]$$

$$= (\cancel{X}^TX)^{-1}X^TX\, E[\beta \mid X] + (X^TX)^{-1}X^T E[\varepsilon]$$

$$= E[\beta \mid X] + 0 = \boxed{\beta}$$

c) By tower property of expectation,

$$\mathrm{Cov}[\hat{\beta}] = E[\mathrm{Cov}[\hat{\beta} \mid X]] + \mathrm{Cov}[E[\hat{\beta} \mid X]]$$

Using result from a),

$$= \sigma^2 E[(X^TX)^{-1}] + \mathrm{Cov}[\beta]$$

d.) Essentially, we repeatedly sample from X (each sample $X_{(i)}$), then

for i in 1 : B {

(1) Calculate $(X_{(i)}^T X_{(i)})^{-1}$ and track moving avg.

(2) Calculate $\hat{\beta}_{(i)}$ via ols

(3) We use (1) to calculate

$\sigma^2 \mathbb{E}[(X^T X)^{-1}]$ by using avg. as an approximation for expectation

(4) We use (2) to calculate

$Cov[\hat{\beta}]$ by approximating $\hat{\beta} = \frac{\sum_{i=1}^{B} \hat{\beta}_{(i)}}{B}$

---

4. a.) With the first approach we only calculate confidence intervals on a pointwise basis. However with the 2nd approach a confidence region is calculated using ~~the~~ all ~~entire~~ parameters, hence requiring more coverage ~~to~~ ensure the same confidence level is met. Hence, the confidence bands produced by 2nd method will be wider

b.) Yes, the bands constructed via approach 2 as $f(x_0)$ is contained within both upper & lower bands $\forall x_0$.

5.

The key idea we exploit is that the expected test error is the same regardless of M. In other words,

$$E\left[\frac{1}{M}\sum_{i=1}^{M}(y_i' - \hat{\beta}^T x_i')^2\right]$$

$$= E\left[\frac{1}{N}\sum_{i=1}^{N}(y_i' - \hat{\beta}^T x_i')^2\right]$$

i.e. WLOG, test & training sets are same size.

Let $\tilde{B}$ be OLS solution for $(x_i', y_i')\ \forall i$, then

$$E\left[\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{\beta}^T x_i)^2\right] = E\left[\frac{1}{N}\sum_{i=1}^{N}(\tilde{y}_i' - \tilde{\beta}^T x_i')^2\right]$$

Since the avg. errors follow same distribution. But as $\tilde{B}$ is OLS, then by definition

$$E\left[\frac{1}{N}\sum_{i=1}^{N}(y_i' - \tilde{\beta}^T x_i')^2\right] \leq E\left[\frac{1}{N}\sum_{i=1}^{N}(y_i' - \hat{\beta}^T x_i')^2\right]$$

$$\therefore \boxed{E[R_{tr}(\hat{\beta})] \leq E[R_{te}(\hat{\beta})]}$$

6 a.) The OLS solution is not unique as when $p >> N$, the matrix $X^T X$ is invertible and thus multiple hyperplanes with RSS = can be fit.

b.) The solution is unique because adding λI_p to $X^T X$ ensures invertibility and by extension a unique solution.

c.) This approach causes the solution to converge to the OLS solution $\hat{\beta}$ that minimizes the 2-norm of the residual vector $\vec{e} = Y - X\beta$

d.) Suppose SVD of X given by

$X = UDV^T$ Then,

$$\hat{\beta} = V(D^2)^{-1} D U^T y$$

$$= \boxed{V D^{-1} U^T y}$$