Ex 2.3 Derive $d(p,N) = \left(1 - \frac{1}{2}^{1/N}\right)^{1/p}$

Treat the distance between data points and the origin as a random variable X, since the data points are equally distributed, the cdf of X is

$$F(X < x) = x^p, \quad x \in [0,1]$$

Then the closest distance from data points to the origin is order statistics, whose cdf is

$$F_1(x) = 1 - (1 - F(x))^N$$

and whose median satisfies $F(d(p,N)) = \frac{1}{2}$, then it is derived.

Ex. 2.5

(a) Derive equation (2.27). The last line makes use of (3.8) through a conditioning argument.

(b) Derive equation (2.28), making use of the *cyclic* property of the trace operator [trace(AB) = trace(BA)], and its linearity (which allows us to interchange the order of trace and expectation).

(a)
$$\begin{aligned}
\mathrm{EPE}(x_0) &= \mathrm{E}_{y_0|x_0}\mathrm{E}_{\mathcal{T}}(y_0 - \hat{y}_0)^2 \\
&= \mathrm{Var}(y_0|x_0) + \mathrm{E}_{\mathcal{T}}[\hat{y}_0 - \mathrm{E}_{\mathcal{T}}\hat{y}_0]^2 + [\mathrm{E}_{\mathcal{T}}\hat{y}_0 - x_0^T\beta]^2 \\
&= \mathrm{Var}(y_0|x_0) + \mathrm{Var}_{\mathcal{T}}(\hat{y}_0) + \mathrm{Bias}^2(\hat{y}_0) \\
&= \sigma^2 + \mathrm{E}_{\mathcal{T}}x_0^T(\mathbf{X}^T\mathbf{X})^{-1}x_0\sigma^2 + 0^2. \qquad (2.27)
\end{aligned}$$

$$= E_{y_0|x_0} E_{\mathcal{T}} (x_0^T\beta + \varepsilon - \hat{y}_0)^2$$

$$= E_{y_0|x_0} E_{\mathcal{T}} [(x_0^T\beta - \hat{y}_0)^2 + 2\varepsilon(x_0^T\beta - \hat{y}_0) + \varepsilon^2]$$

$$= E_{\mathcal{T}}[(x_0^T\beta - \hat{y}_0)^2] + 2\underbrace{E_{y_0|x_0}}_{E_{y_0|x_0}(\varepsilon) = 0}[\varepsilon(x_0^T\beta - \hat{y}_0)] + E_{y_0|x_0}(\varepsilon^2)$$

$$= E_{y_0|x_0}(\varepsilon^2) + E_{\mathcal{T}}[x_0^T\beta - E(\hat{y}_0) + E(\hat{y}_0) - \hat{y}_0]^2$$

(b) $x_0\ (p\times 1)$ $X\ (N\times p)$ $\mathrm{Cov}(X)^{-1} = (X^TX)^{-1}$

$$\begin{aligned}
\mathrm{E}_{x_0}\mathrm{EPE}(x_0) &\sim \mathrm{E}_{x_0} \underset{1\times p}{x_0^T} \underset{p\times p}{\mathrm{Cov}(X)^{-1}} \underset{p\times 1}{x_0} \sigma^2/N + \sigma^2 \quad (\to 1\times 1) \\
&= \mathrm{trace}[\mathrm{Cov}(X)^{-1}\mathrm{Cov}(x_0)]\sigma^2/N + \sigma^2 \\
&= \sigma^2(p/N) + \sigma^2. \qquad (2.28)
\end{aligned}$$

Since trace(AB) = trace(BA)

$$= \mathrm{trace}(E_{x_0}(x_0^T \mathrm{Cov}(X)^{-1} x_0))\, \sigma^2/N + \sigma^2$$

$$= \mathrm{trace}(E_{x_0}(x_0 x_0^T \mathrm{Cov}(X)^{-1}))\, \sigma^2/N + \sigma^2$$

$$= \mathrm{trace}(E_{x_0}[\underbrace{x_0 x_0^T}_{\sigma^2 I_p} \underbrace{(X^TX)^{-1}}_{\sigma^{-2} I_p}])\, \sigma^2/N + \sigma^2$$

$$= \sigma^2\sigma^{-2}\,\mathrm{trace}(I_p)\cdot \sigma^2/N + \sigma^2$$

$$= \sigma^2(p/N) + \sigma^2$$

Ex 2.6

Ex. 2.6 Consider a regression problem with inputs x_i and outputs y_i, and a parameterized model $f_\theta(x)$ to be fit by least squares. Show that if there are observations with *tied* or *identical* values of x, then the fit can be obtained from a reduced weighted least squares problem.

Ex. 2.6 OLS: $\hat{f}_\theta(x) = X(X^TX)^{-1}X^TY$ (1)

for example, , $X_u = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$ while $X = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \to x_3 = x_1 + x_2$

where we derived the weight $W = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 1 \end{bmatrix}$

Then $X = WX_u = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 1 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} x_1 \\ x_2 \\ x_1 + x_2 \end{bmatrix} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$

Plug it in (1): $\hat{f}_\theta(X) = WX_u(X_u^TW^TWX_u)^{-1}X_u^TW^TY$

which is a weighted LS fit.

Obviously, the size of X_u is smaller than that of X, $\hat{f}_\theta(X)$ is a reduced weighted LS.