Exercise 4: An Alternative to Least Squares? (3 pts)

Suppose we are in a setting with a dataset $X \in \mathbb{R}^{n\times d}$, and labels are generated according to $Y = XW + \varepsilon$, where $W \in \mathbb{R}^d$, $Y \in \mathbb{R}^n$, and $\varepsilon \in \mathbb{R}^n$ is a random vector, where all entries are independent random variables with 0 mean and variance σ^2 (you can imagine Gaussian, but it isn't necessary). As we saw in class, the least-squares solution $\hat{W}$ can be written as $\hat{W} = (X^TX)^{-1}X^TY$ -- this is a linear transformation of the response vector Y. Consider some *different* linear transformation $\left((X^TX)^{-1}X^T + N\right)Y$, where $N \in \mathbb{R}^{d\times n}$ is a non-zero matrix.

1. (1 pt) Show that the expected value of this linear transformation is $(I_d + NX)W$. Conclude that its expected value is W if and only if $NX = 0$. (1 pt)
   Ans: **Answer 4.1**
2. (2 pts) Compute the covariance matrix of this linear transformation when $NX = 0$, and show that it is equal to $\sigma^2(X^TX)^{-1} + \sigma^2NN^T$. Since the former term is the covariance of the least squares solution[a] and the latter matrix is positive semi-definite, this implies that this alternative estimator only increased the variance of our estimate.
   Ans: **Answer 4.2**

[a] Verify this for yourself, but no need to submit it.

Answer 4.1

Let us name this linear transformation as $\hat{W}$.

$$
\begin{aligned}
E[\hat{W}] &= E\left[\left((X^TX)^{-1}X^T + N\right)Y\right] && (63)\\
&= E\left[\left((X^TX)^{-1}X^T + N\right)(XW+\varepsilon)\right] && (64)\\
&= E\left[\left((X^TX)^{-1}X^T + N\right)XW + \left((X^TX)^{-1}X^T + N\right)\varepsilon\right] && (65)\\
&= E\left[\left((X^TX)^{-1}X^T + N\right)XW\right] + E\left[\left((X^TX)^{-1}X^T + N\right)\varepsilon\right] && (66)\\
&= E\left[\left(\cancelto{I_d}{(X^TX)^{-1}X^TX} + NX\right)W\right] + E\left[\left((X^TX)^{-1}X^T + N\right)\right]\cancelto{0}{E[\varepsilon]} && (67)\\
&= E\left[(I_d + NX)W\right] && (68)\\
&= (I_d + NX)W && (69)
\end{aligned}
$$

We may also prove the expected value is W if and only if $NX = 0$:

$$
\begin{aligned}
E[\hat{W}] &= (I_d + NX)W && (70)\\
&= W + NXW && (71)\\
E[\hat{W}] - W &= NXW && (72)\\
&\text{If } NXW \neq 0 \Rightarrow E[\hat{W}] - W \neq 0 \Rightarrow E[\hat{W}] \neq W && (73)\\
&\text{If } NXW = 0 \Rightarrow E[\hat{W}] - W = 0 \Rightarrow E[\hat{W}] = W && (74)\\
&\therefore E[\hat{W}] = W \ \textbf{iff}\ NXW = 0 && (75)
\end{aligned}
$$

Q.E.D.

Answer 4.2

Similar to Answer 4.1, we may derive $\hat{W}$ and $(\hat{W} - W)$:

$$
\begin{aligned}
\because \quad \hat{W} &= \left((X^TX)^{-1}X^T + N\right)XW + \left((X^TX)^{-1}X^T + N\right)\varepsilon && (76)\\
&= \overset{I_d}{\cancel{(X^TX)^{-1}X^TX}}W + \overset{0}{\cancel{NX}}W + (X^TX)^{-1}X^T\varepsilon + N\varepsilon && (77)\\
&= W + (X^TX)^{-1}X^T\varepsilon + N\varepsilon && (78)\\
\therefore \hat{W} - W &= (X^TX)^{-1}X^T\varepsilon + N\varepsilon && (79)
\end{aligned}
$$

We may now compute the covariance matrix:

$$
\begin{aligned}
\Sigma &= E\left[(\hat{W} - W)(\hat{W} - W)^T\right] && (80)\\
&= E\left[\left((X^TX)^{-1}X^T\varepsilon + N\varepsilon\right)\left((X^TX)^{-1}X^T\varepsilon + N\varepsilon\right)^T\right] && (81)\\
&= E\left[\left((X^TX)^{-1}X^T\varepsilon + N\varepsilon\right)\left(\varepsilon^TX(X^TX)^{-1} + \varepsilon^TN^T\right)\right] && (82)\\
&= E\left[(X^TX)^{-1}X^T\varepsilon\varepsilon^TX(X^TX)^{-1} + N\varepsilon\varepsilon^TX(X^TX)^{-1} + \varepsilon^TN^T(X^TX)^{-1}X^T + N\varepsilon\varepsilon^TN^T\right] && (83)\\
&= E\left[(X^TX)^{-1}(X^TX)\varepsilon\varepsilon^T(X^TX)^{-1} + NX\varepsilon\varepsilon^T(X^TX)^{-1} + N^TX^T\varepsilon^T(X^TX)^{-1} + N\varepsilon\varepsilon^TN^T\right] && (84)\\
&= E\left[\overset{I_d}{\cancel{(X^TX)^{-1}(X^TX)}}\varepsilon\varepsilon^T(X^TX)^{-1} + \overset{0}{\cancel{NX}}\varepsilon\varepsilon^T(X^TX)^{-1} + \overset{0}{\cancel{(NX)^T}}\varepsilon^T(X^TX)^{-1} + N\varepsilon\varepsilon^TN^T\right] && (85)\\
&= E\left[\varepsilon\varepsilon^T(X^TX)^{-1} + N\varepsilon\varepsilon^TN^T\right] && (86)\\
&= E\left[\varepsilon\varepsilon^T(X^TX)^{-1}\right] + E\left[N\varepsilon\varepsilon^TN^T\right] && (87)\\
&= E\left[\varepsilon\varepsilon^T\right](X^TX)^{-1} + E\left[\varepsilon\varepsilon^T\right]NN^T && (88)\\
&= \sigma^2(X^TX)^{-1} + \sigma^2NN^T && (89)
\end{aligned}
$$

Q.E.D.

Remark 4.2: ε Term

$E[\varepsilon\varepsilon^T] = \sigma^2 I_d$: The covariance of additive noise term is the variance, which is σ^2 as provided
$E[\varepsilon] = 0$: The expectation of the additive noise term is the mean, which is 0 as provided