ST5227 Assignment 1.

Qi Shuoli (A0274285W) E1124031

Linear Regression

Problem 1.

Since we let $X_{new} = XD$ for some diagonal matrix D

So $Y = XD\beta + \varepsilon$ and loss function $L(\beta; Y, n) = (Y - XD\beta)^T(Y - XD\beta)$

Compute the first and second derivatives: $\frac{\partial}{\partial \beta} L(\beta; Y, XD) = -2DX^T(Y - XD\hat{\beta}) = -2DX^T(Y - XD\beta)$

$(\frac{\partial^2}{\partial \beta \partial \beta^T} = 2DX^T(XD))$

Let $\frac{\partial}{\partial \beta} L = 0$ so $-2DX^TY + 2DX^TXD\beta = 0 \Rightarrow 2DX^TXD\beta = 2DX^TY$

so $\hat{\beta}_{new} = (DX^TXD)^{-1}DX^TY = D^{-1}(X^TX)^{-1}X^TY = D^{-1}\hat{\beta}^{ols}$

comparing $\hat{\beta}^{ols} = (X^TX)^{-1}X^TY$. $\hat{\beta}_{new}$ is having another D^{-1}.

Consequently, standardizing the predictors is not necessary before ordinary least square.

The $X_{new}\tilde{\beta}$ is same as original, but just with $\hat{\beta}^{ols}$ shrink to D^{-1}. So the final result is not affected. The other property of the result is actually not affected.

Problem 2.

Let $\hat{\beta}^{new} = CY$. since $Y = X\beta + \varepsilon$ that is $\hat{\beta}^{new} = C(X\beta + \varepsilon) = CX\beta + C\varepsilon$

Since $\hat{\beta}^{new}$ is an unbiased estimate of β. where C is a $p+1 \times n$ matrix

So $E(\hat{\beta}^{new}) = \beta = E(CY)$

So $\hat{\beta}^{new}X = Y$. $CYX = Y$ so $CX = I$, CX equal the identity matrix

Consequently. $\hat{\beta}^{new} = CX\beta + C\varepsilon = \beta + C\varepsilon$.

$Var(\hat{\beta}^{new}) = var(CY) = C var(Y) C^T = \sigma^2 CC^T$

$= \sigma^2 (C + (X^TX)^{-1}X^T - (X^TX)^{-1}X^T)(C + (X^TX)^{-1}X^T - (X^TX)^{-1}X^T)^T$

$= \sigma^2 [(X^TX)^{-1}X^TX(X^TX)^{-1} + (X^TX)^{-1}X^T[C - (X^TX)^{-1}X]^T + [C - (X^TX)^{-1}X^T]X(X^TX)^{-1} + [C - (X^TX)^{-1}X^T][C - (X^TX)^{-1}X^T]^T]$

$= \sigma^2 [(X^TX)^{-1} + (X^TX)^{-1}(C - (X^TX)^{-1}X^T)^TX + (C - (X^TX)^{-1}X^T)X(X^TX)^{-1} + (C - (X^TX)^{-1}X^T)(C - (X^TX)^{-1}X^T)^T]$

Additionally $(C-(X^TX)^{-1}X^T)X = CX - I = 0$

So $Var(\hat{\beta}^{new}) = \sigma^2[(X^TX)^{-1}+(C-(X^TX)^{-1}X^T)(C-(X^TX)^{-1}X^T)^T]$

$= var(\hat{\beta}^{ols}) + \sigma^2(C-(X^TX)^{-1}X^T)(C-(X^TX)^{-1}X^T)^T$

Since for any matrix A, each term of diagonal of AA^T is greater than or equal to 0.

So $C-(X^TX)^{-1}X^T$ is semi-definite,

so $Var(\hat{\beta}^{new}-\hat{\beta}^{ols}) = \sigma^2(C-(X^TX)^{-1}X^T)(C-(X^TX)^{-1}X^T)^T \geq 0 \Rightarrow var(\hat{\beta}^{new}) \geq var(\hat{\beta}^{ols})$

Extension to homoskedastic data

Problem 3

(1). The loss function $L(\beta) = (Y-X\beta)^T W (Y-X\beta)$

$\hat{\beta}^{wls} = \arg\min_\beta L(\beta)$ and get the gradient of $L(\beta)$ with respect to β to equal to 0 to get $\hat{\beta}^{wls}$:

$\frac{\partial}{\partial\beta}L(\beta) = 0 \Rightarrow -2X^TW(Y-X\beta) = 0 \Rightarrow X^TWY = X^TWX\beta$

$\therefore \hat{\beta}^{wls} = (X^TWX)^{-1}X^TWY$

Then, $\frac{\partial^2}{\partial\beta\partial\beta^T}L(\beta) = 2X^TWX \geq 0$, so there is the minimum value for $L(\beta)$ when $\hat{\beta}^{wls} = (X^TWX)^{-1}X^TWY$

(2). $Bias(\hat{\beta}^{wls}) = E(\hat{\beta}^{wls}) - \beta$

$E(\hat{\beta}^{wls}) = (X^TWX)^{-1}X^TWE(Y) = (X^TWX)^{-1}X^TW(X\beta+E(\varepsilon))$

$E(\varepsilon) = 0. \quad \therefore E(\beta^{wls}) = (X^TWX)^{-1}X^TWX\beta = \beta$

so $Bias(\hat{\beta}^{wls}) = \beta - \beta = 0$

(3). $var(\hat{\beta}^{wls}) = var[(X^TWX)^{-1}X^TWY] = (X^TWX)^{-1}X^TW\,Var(Y)\,WX((X^TWX)^{-1})^T$

Since W is a diagonal matrix with weights $W = \frac{1}{\sigma^2}$ on diagnol.

Diagonal matrices are symmetric, so $W = W^T$

$((X^TWX)^{-1})^T = ((X^TWX)^T)^{-1} = (XW^TX^T)^{-1} = (XWX^T)^{-1}$, so XWX^T also symmetric.

So $var(\hat{\beta}^{wls}) = (X^TWX)^{-1}X^TW\,Var(Y)\,WX(X^TWX)^{-1}$

For $Var(Y) = var(X\beta+\varepsilon) = var(X\beta) + var(\varepsilon) + 2cov(X\beta, \varepsilon)$

$X\beta$ is not Random, error ε are independent of predictor X, so

We do have $Var(X\beta) = 0$, $cov(X\beta, \varepsilon) = 0$. So $var(X\beta) = 0$, $cov(X\beta, \varepsilon) = 0$.

So $var(Y) = var(\varepsilon) = W^{-1}$

$var(\hat{\beta}^{wls}) = (X^TWX)^{-1}X^TWW^{-1}WX(X^TWX)^{-1} = (X^TWX)^{-1}X^TWX(X^TWX)^{-1} = (X^TWX)^{-1}$

Problem 4.

$\hat{\beta}^{wls} = (X^TWX)^{-1}X^TWY$ $E(\hat{\beta}^{wls}) = \beta$ $var(\hat{\beta}^{wls}) = (X^TWX)^{-1}$

Let $\hat{\beta}^{new} = CY$. We want to show $var(\hat{\beta}^{new}) \geq var(\hat{\beta}^{wls})$ (β unbiased)

Since $Y = X\beta + \varepsilon$. $\hat{\beta}^{new} = C(X\beta+\varepsilon) = CX\beta + C\varepsilon = \beta + C\varepsilon$

$var(\hat{\beta}^{new}) = var(C\varepsilon + \beta)$. since β represent true coefficients that aim to estimate.

Coefficient does not change with the error term ε or other random variable. So β is considered a constant. So $Var(\hat{\beta}^{new}) = var(C\varepsilon) = CW^{-1}C^T$

$$
\begin{aligned}
Var(\hat{\beta}^{new}) = CW^{-1}C^T &= (C + (X^TX)^{-1}X^T - (X^TX)^{-1}X^T)W^{-1}(C + (X^TX)^{-1}X^T - (X^TX)^{-1}X^T)^T \\
&= W^{-1}\left[(X^TX)^{-1}X^TX(X^TX)^{-1} + (X^TX)^{-1}X^T[C - (X^TX)^{-1}X^T] + [C - (X^TX)^{-1}X^T][C - (X^TX)^{-1}X^T]^T\right] \\
&= W^{-1}\left[(X^TX)^{-1} + (X^TX)^{-1}(C - (X^TX)^{-1}X^T)^TX + (C - (X^TX)^{-1}X^T)X(X^TX)^{-1} + (C - (X^TX)^{-1}X^T)(C - (X^TX)^{-1}X^T)^T\right]
\end{aligned}
$$

Additionally. $W^{-1}(C - (X^TX)^{-1}X^T)^TX = CW^{-1}X - I = CX - I = 0$

$$
\begin{aligned}
\text{so } var(\hat{\beta}^{new}) &= (X^TWX)^{-1} + W^{-1}\left[(C - (X^TX)^{-1}X^T)(C - (X^TX)^{-1}X^T)^T\right] \\
&= var(\hat{\beta}^{wls}) + \left[W^{-\frac{1}{2}}(C - (X^TX)^{-1}X^T)(C - (X^TX)^{-1}X^T)W^{-\frac{1}{2}}\right]^T
\end{aligned}
$$

Since for any matrix A, each term of diagonal of AA^T is greater than or equal to 0.

So $(C - (X^TX)^{-1}X^TW^{-\frac{1}{2}})$ is semi-definite,

so $Var(\hat{\beta}^{new} - \hat{\beta}^{ols}) = W^{-\frac{1}{2}}(C - (X^TX)^{-1}X)(C - (X^TX)^{-1}X)^T \geq 0 \Rightarrow var(\hat{\beta}^{new}) \geq var(\hat{\beta}^{ols})$

Problem 5

$$E[(Y - \hat{Y})^2] = E[E(Y - \hat{f}(x))^2 \mid \hat{f}(x)] = E[(f(x) - \hat{f}(x))^2] + var(\varepsilon)$$

$$= E[f^2(x) - 2f(x)\hat{f}(x) + \hat{f}(x)^2] + Var(\varepsilon)$$

$$= E(f^2(x)) - E(2f(x)\hat{f}(x)) + E(\hat{f}(x)^2) + Var(\varepsilon)$$

$$= E[f^2(x)] - 2f(x)E[\hat{f}(x)] + f^2(x) + Var(\varepsilon)$$

$$= E[\hat{f}^2(x)] - [E(\hat{f}(x))]^2 + [E(\hat{f}(x))]^2 - 2f(x)E[\hat{f}(x)] + f^2(x) + Var(\varepsilon)$$

$$= E[\hat{f}(x)^2] - 2E[\hat{f}(x)]^2 + E[\hat{f}(x)]^2 + [E(\hat{f}(x)) - f(x)]^2 + Var(\varepsilon)$$

$$= E[\hat{f}(x)^2] - 2E[\hat{f}(x)]E[\hat{f}(x)] + E[E\hat{f}(x)]^2 + [E(\hat{f}(x)) - f(x)]^2 + Var(\varepsilon)$$

$$= [E[\hat{f}(x)] - E[E(\hat{f}(x))]]^2 + [E(\hat{f}(x)) - f(x)]^2 + Var(\varepsilon)$$

$$= E[\hat{f}(x) - E(\hat{f}(x))]^2 + [E(\hat{f}(x)) - f(x)]^2 + Var(\varepsilon)$$

$$= Var(\hat{f}(x)) + (Bias(\hat{f}(x)))^2 + Var(\varepsilon)$$

Problem b

Let loss function $L(\beta) = (Y - X\beta)^T(Y - X\beta) + \lambda\beta^T\beta$

so $\hat{\beta}^{ridge} = \underset{\beta}{\operatorname{argmin}} L(\beta)$ and then use gradient to find $\hat{\beta}^{ridge}$

$$\frac{\partial}{\partial\beta}(Y - X\beta)^T(Y - X\beta) + \lambda\beta^T\beta = -2X^T(Y - X\beta) + 2\lambda\beta = 0$$

$$\therefore -2X^T(Y - X\beta) + 2\lambda\beta = 0 \Rightarrow X^T(Y - X\beta) = \lambda\beta \Rightarrow X^TY - X^TX\beta = \lambda\beta$$

$$\therefore X^TX\beta + \lambda\beta = X^TY \Rightarrow (X^TX + \lambda I)\beta = X^TY$$

$$\therefore \hat{\beta}^{ridge} = (X^TX + \lambda I)^{-1}X^TY$$

$\frac{\partial}{\partial\beta\partial\beta^T}L(\beta) = 2X^TX + 2\lambda I \geq 0$ so there is the minimum value for $L(\beta)$ when $\hat{\beta}^{ridge} = (X^TX + \lambda I)^{-1}X^TY$

Problem 7.

$$\hat{\beta}^{ridge} = (X^TX + \lambda I)^{-1}X^TY = (X^TX + \lambda I)^{-1}X^T(X\beta + \varepsilon)$$

$$= (X^TX + \lambda I)^{-1}X^TX\beta + (X^TX + \lambda I)^{-1}X^T\varepsilon$$

$$E(\hat{\beta}^{ridge}) = E[(X^TX + \lambda I)^{-1}X^TY] = E[(X^TX + \lambda I)^{-1}X^T(X\beta + \varepsilon)]$$

$$= (X^TX + \lambda I)^{-1}X^TX\beta$$

So $\text{Bias}(\hat{\beta}^{ridge}) = E(\hat{\beta}^{ridge}) - \beta = (X^TX+\lambda I)^{-1}X^TX\beta - \beta = \left[(X^TX+\lambda I)^{-1}X^TX - I\right]\beta$

$= (X^TX+\lambda I)^{-1}\left[X^TX - (X^TX+\lambda I)\right]\beta = -\lambda(X^TX+\lambda I)^{-1}\beta$

$\text{Var}(\hat{\beta}^{ridge}) = (X^TX+\lambda I)^{-1}X^T\text{Var}(\varepsilon)X(X^TX+\lambda I)^{-1}$

Since $\text{Var}(\varepsilon) = \sigma^2 I$, $\text{Var}(\hat{\beta}^{ridge}) = \sigma^2(X^TX+\lambda I)^{-1}X^TX(X^TX+\lambda I)^{-1}$

Problem 8.

$\hat{\beta}^{ols} = (X^TX)^{-1}X^TY \quad \hat{\beta}^{ridge} = (X^TX+\lambda I)^{-1}X^TY$

$\text{Var}(\hat{\beta}^{ols}) = (X^TX)^{-1}X^T\sigma^2 I X(X^TX)^{-1} = (X^TX)^{-1}\sigma^2$

$\text{Var}(\hat{\beta}^{ridge}) = (X^TX+\lambda I)^{-1}X^TX(X^TX+\lambda I)^{-1}\sigma^2$

$\text{Var}(\hat{\beta}) - \text{Var}(\hat{\beta}^{ridge}) = \sigma^2(X^TX+\lambda I)^{-1}(2\lambda I + \lambda^2(X^TX)^{-1})(X^TX+\lambda I)^{-1}$

$(X^TX)^{-1}$ is the inverse of a positive definite matrix so $(X^TX)^{-1}$ is positive semidefinite $\left(\begin{matrix} y=Ax \\ y^TA^{-1}y = x^TA^TA^{-1}Ax \\ = x^TAx \geq 0 \end{matrix}\right)$

So at the same time $2\lambda I + \lambda^2(X^TX)^{-1}$ should be positive semidefinite. $(x^T(A+B)x = x^TAx + x^TBx \geq 0)$

Consequently $\text{Var}(\hat{\beta}) - \text{Var}(\hat{\beta}^{ridge}) = \sigma^2(X^TX+\lambda I)^{-1}(2\lambda I + \lambda^2(X^TX)^{-1})(X^TX+\lambda I)^{-1} \geq 0$

So $\text{Var}(\hat{\beta}) \geq \text{var}(\hat{\beta}^{ridge})$

Problem 9.

$\hat{\beta}^{ridge} = \underset{\beta}{\text{argmin}}(Y-X\beta)^T(Y-X\beta) + \lambda\beta^T\beta$

$X_{new} = XD$

$\hat{\beta}^{ridge}_{new} = \underset{\beta}{\text{argmin}}\left\{(Y-XD\beta)^T(Y-XD\beta) + \lambda\beta^T\beta\right\}$

$\frac{\partial}{\partial\beta}(Y-XD\beta)^T(Y-XD\beta) + \lambda\beta^T\beta = 2XD^T(Y-XD\beta) + 2\lambda\beta = -2X^TD^T(Y-XD\beta) + 2\lambda\beta = 0$

$(XD)^TY - (XD)^TXD\beta = \lambda\beta \Rightarrow (XD)^TXD\beta + \lambda\beta = (XD)^TY \Rightarrow ((XD)^TXD + \lambda I)\beta = (XD)^TY$

$\therefore \hat{\beta}^{ridge}_{new} = \left[(XD)^TXD + \lambda I\right]^{-1}(XD)^TY = D^{-1}(X^TX + \lambda D^{-2})^{-1}X^TY$ compare $\hat{\beta}^{ridge} = (X^TX+\lambda I)^{-1}X^TY$

Equation be the same if and only $D^{-2} = I$.

As ridge regression not linear transform, so we need to standardize first.