回归分析 Homework 02

Due: Dec. 19, 2024
姓名: 雍崔扬
学号: 21307140051

Problem 1

The coeffcients β of a linear regression model, $Y = X\beta + \varepsilon$, are estimated by $\hat{\beta} = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y$.
Then the associated fitted values are given by $\hat{Y} = X\hat{\beta} = X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}Y = HY$,
where $H = X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}$ is the hat matrix, which is a projection operator.
Hence, linear regression projects the response Y onto $\operatorname{span}(X)$ (i.e., column space of design matrix)
Consequently, the residuals $\hat{\varepsilon}$ and $\hat{Y}$ are orthogonal.

Now consider the ridge estimator of the regression coeffcients: $\hat{\beta}(\lambda) = (X^{\mathrm{T}}X + \lambda I)^{-1}X^{\mathrm{T}}y$.
Let $\hat{Y}(\lambda) = X\hat{\beta}(\lambda)$ be the vector of associated fitted values.

- ① Show that the matrix $H(\lambda) = X(X^{\mathrm{T}}X + \lambda I)^{-1}X^{\mathrm{T}}$ is not a projection matrix (for any $\lambda > 0$)
- ② Show that the ridge fit $\hat{Y}(\lambda)$ is not orthogonal to the associated ridge residuals $\hat{\varepsilon}(\lambda)$ (for any $\lambda > 0$)
- ③ Given that $\varepsilon \sim N(0_n, \sigma^2 I_n)$, derive the distribution of ridge residuals $\hat{\varepsilon}(\lambda)$

Part (1)

Show that the matrix $H(\lambda) = X(X^{\mathrm{T}}X + \lambda I)^{-1}X^{\mathrm{T}}$ is not a projection matrix (for any $\lambda > 0$)

Proof:
根据泛函分析中的结论:
Hilbert 空间 $(V, \langle\cdot,\cdot\rangle)$ 中的线性算子 H 是投影算子当且仅当 H 是幂等且自伴的，即满足 $\begin{cases} H^2 = H \\ H^* = H \end{cases}$
特别地，有限维内积空间一定是 Hilbert 空间.
复 Euclid 空间上的伴随算子 H^* 的表示矩阵即为算子 H 表示矩阵的共轭转置.

考虑到 $H(\lambda) = X(X^{\mathrm{T}}X + \lambda I)^{-1}X^{\mathrm{T}}$ 一定是对称阵,
故我们要说明它不是投影算子，只需说明它不是幂等算子即可,
即不满足 $(H(\lambda))^2 = H(\lambda)$，也即其存在某个特征值不是 0 或 1.

设 $X \in \mathbb{R}^{n\times(p+1)}$ 的奇异值分解为 $X = U\Sigma V^{\mathrm{T}}$
其中 $U \in \mathbb{R}^{n\times n}, V \in \mathbb{R}^{(p+1)\times(p+1)}$ 是实正交阵,
而 $\Sigma \in \mathbb{R}^{n\times(p+1)}$ 的对角元 $\sigma_1, \ldots, \sigma_{p+1}$ 均为正实数 (因为模型的列满秩假设规定 $\operatorname{rank}(X) = p+1 < n$)
则对于任意给定的 $\lambda > 0$，我们都有:

$$
\begin{aligned}
H(\lambda) &= X(X^{\mathrm{T}}X+\lambda I)^{-1}X^{\mathrm{T}} \\
&= U\Sigma V^{\mathrm{T}}[(U\Sigma V^{\mathrm{T}})^{\mathrm{T}}(U\Sigma V^{\mathrm{T}})+\lambda I]^{-1}(U\Sigma V^{\mathrm{T}})^{\mathrm{T}} \\
&= U\Sigma V^{\mathrm{T}}(V\Sigma^{\mathrm{T}}\Sigma V^{\mathrm{T}}+\lambda I)^{-1}V\Sigma^{\mathrm{T}}U^{\mathrm{T}} \\
&= U\Sigma(\Sigma^{\mathrm{T}}\Sigma+\lambda I)^{-1}\Sigma^{\mathrm{T}}U^{\mathrm{T}} \\
&= U\begin{bmatrix} \frac{\sigma_1^2}{\sigma_1^2+\lambda} & & & & & \\ & \ddots & & & & \\ & & \frac{\sigma_{p+1}^2}{\sigma_{p+1}^2+\lambda} & & & \\ & & & 0 & & \\ & & & & \ddots & \\ & & & & & 0 \end{bmatrix}U^{\mathrm{T}} \\
&= U\mathrm{diag}\left\{\frac{\sigma_1^2}{\sigma_1^2+\lambda},\dots,\frac{\sigma_{p+1}^2}{\sigma_{p+1}^2+\lambda},\underbrace{0,\dots,0}_{n-p-1}\right\}U^{\mathrm{T}}
\end{aligned}
$$

$$
(H(\lambda))^2 = U\mathrm{diag}\left\{\left(\frac{\sigma_1^2}{\sigma_1^2+\lambda}\right)^2,\dots,\left(\frac{\sigma_{p+1}^2}{\sigma_{p+1}^2+\lambda}\right)^2,\underbrace{0,\dots,0}_{n-p-1}\right\}U^{\mathrm{T}}
$$

显然对于任意 $\lambda>0$，$H(\lambda)$ 都有 $p+1$ 个特征值 $\frac{\sigma_1^2}{\sigma_1^2+\lambda},\dots,\frac{\sigma_{p+1}^2}{\sigma_{p+1}^2+\lambda}$ 既不为 0，又不为 1
因此 $H(\lambda)$ 不是幂等算子 (即不满足 $(H(\lambda))^2=H(\lambda)$)
具体来说:

$$
\begin{aligned}
&H(\lambda)-(H(\lambda))^2 \\
&= U\mathrm{diag}\left\{\frac{\sigma_1^2}{\sigma_1^2+\lambda},\dots,\frac{\sigma_{p+1}^2}{\sigma_{p+1}^2+\lambda},\underbrace{0,\dots,0}_{n-p-1}\right\}U^{\mathrm{T}} - U\mathrm{diag}\left\{\left(\frac{\sigma_1^2}{\sigma_1^2+\lambda}\right)^2,\dots,\left(\frac{\sigma_{p+1}^2}{\sigma_{p+1}^2+\lambda}\right)^2,\underbrace{0,\dots,0}_{n-p-1}\right\}U^{\mathrm{T}} \\
&= U\mathrm{diag}\left\{\frac{\lambda\sigma_1^2}{(\sigma_1^2+\lambda)^2},\dots,\frac{\lambda\sigma_{p+1}^2}{(\sigma_{p+1}^2+\lambda)^2},\underbrace{0,\dots,0}_{n-p-1}\right\}U^{\mathrm{T}} \\
&\neq 0_{n\times n}
\end{aligned}
$$

Part (2)

Show that the ridge fit $\hat{y}(\lambda)$ is not orthogonal to the associated ridge residuals $\hat{\varepsilon}(\lambda)$ (for any $\lambda>0$)

Proof:
对于任意 $\lambda>0$，我们都有:

$$
\begin{aligned}
(\hat{y}(\lambda))^{\mathrm{T}}\hat{\varepsilon}(\lambda) &= (\hat{H}(\lambda)y)^{\mathrm{T}}(y-\hat{H}(\lambda)y) \quad (\text{note that } (\hat{H}(\lambda))^{\mathrm{T}}=\hat{H}(\lambda)) \\
&= y^{\mathrm{T}}\hat{H}(\lambda)(I-\hat{H}(\lambda))y \\
&= y^{\mathrm{T}}(\hat{H}(\lambda)-(\hat{H}(\lambda))^2)y \\
&= y^{\mathrm{T}}U\mathrm{diag}\left\{\frac{\lambda\sigma_1^2}{(\sigma_1^2+\lambda)^2},\dots,\frac{\lambda\sigma_{p+1}^2}{(\sigma_{p+1}^2+\lambda)^2},\underbrace{0,\dots,0}_{n-p-1}\right\}U^{\mathrm{T}}y \\
&= (U^{\mathrm{T}}y)^{\mathrm{T}}\mathrm{diag}\left\{\frac{\lambda\sigma_1^2}{(\sigma_1^2+\lambda)^2},\dots,\frac{\lambda\sigma_{p+1}^2}{(\sigma_{p+1}^2+\lambda)^2},\underbrace{0,\dots,0}_{n-p-1}\right\}(U^{\mathrm{T}}y)
\end{aligned}
$$

记正交阵 U 的列向量组为 $U=[u_1,\dots,u_n]\in\mathbb{R}^{n\times n}$
显然 $(\hat{y}(\lambda))^{\mathrm{T}}\hat{\varepsilon}(\lambda)=0$ 当且仅当 $U^{\mathrm{T}}y=[u_1^{\mathrm{T}}y,\dots,u_n^{\mathrm{T}}y]^{\mathrm{T}}$ 的前 $p+1$ 个分量全为零,
即当且仅当 $y\in\mathrm{span}\{u_1,\dots,u_{p+1}\}^{\perp}=\mathrm{span}\{u_{p+1},\dots,u_n\}$
即当且仅当 y 正交于 $\mathrm{span}(X)=\mathrm{span}\{u_1,\dots,u_{p+1}\}$

通常来说，响应变量的观测向量 y 不完全正交于设计矩阵 X 列空间 $\mathrm{span}(X)$,
因此我们可以认为 $(\hat{Y}(\lambda))^{\mathrm{T}}\hat{\varepsilon}(\lambda)\neq 0$ $(\forall\,\lambda>0)$
即对于任意 $\lambda>0$，$\hat{Y}(\lambda)$ 和 $\hat{\varepsilon}(\lambda)$ 都不是正交的.

Part (3)

Given that $\varepsilon \sim N(0_n, \sigma^2 I_n)$, derive the distribution of ridge residuals $\hat{\varepsilon}(\lambda)$

Solution:
对于任意 $\lambda > 0$，我们都有:

$$\begin{aligned}
\hat{\varepsilon}(\lambda) &= y - \hat{y}(\lambda) \\
&= y - H(\lambda)y \\
&= (I - H(\lambda))y \\
&= (I - H(\lambda))(X\beta + \varepsilon)
\end{aligned}$$

根据 $\varepsilon \sim N(0_n, \sigma^2 I_n)$ 可知 $\hat{\varepsilon}(\lambda)$ 也服从多元正态分布，
且其均值向量和协方差矩阵为:

$$\begin{aligned}
\mathrm{E}[\hat{\varepsilon}(\lambda)] &= \mathrm{E}[(I - H(\lambda))(X\beta + \varepsilon)] \\
&= (I - H(\lambda))X\beta + (I - H(\lambda))\mathrm{E}[\varepsilon] \\
&= (I - H(\lambda))X\beta + (I - H(\lambda)) \cdot 0_n \\
&= (I - H(\lambda))X\beta
\end{aligned}$$

$$\begin{aligned}
\mathrm{Cov}(\hat{\varepsilon}(\lambda)) &= \mathrm{Cov}[(I - H(\lambda))(X\beta + \varepsilon)] \\
&= (I - H(\lambda)) \cdot \mathrm{Cov}(\varepsilon) \cdot [(I - H(\lambda))]^{\mathrm{T}} \quad (\text{note that } (H(\lambda))^{\mathrm{T}} = H(\lambda)) \\
&= (I - H(\lambda)) \cdot \sigma^2 I_n \cdot (I - H(\lambda)) \\
&= \sigma^2 (I - H(\lambda))^2
\end{aligned}$$

因此 $\hat{\varepsilon}(\lambda) \sim N((I - H(\lambda))X\beta, \sigma^2 (I - H(\lambda))^2)$

Problem 2

Recall that there exists a $\lambda > 0$ such that $\mathrm{MSE}(\hat{\beta}) > \mathrm{MSE}(\hat{\beta}(\lambda))$
(where $\hat{\beta} = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y$ and $\hat{\beta}(\lambda) = (X^{\mathrm{T}}X + \lambda I)^{-1}X^{\mathrm{T}}y$)
Verify that this carries to the fitted value, that is, there exists a $\lambda > 0$ such that $\mathrm{MSE}(X\hat{\beta}) > \mathrm{MSE}(X\hat{\beta}(\lambda))$

Proof:
考虑 $X\hat{\beta}(\lambda) = H(\lambda)y$ 的均方误差:
(其中 $H(\lambda) = X(X^{\mathrm{T}}X + \lambda I)^{-1}X^{\mathrm{T}}$)

$$\begin{aligned}
\mathrm{MSE}[X\hat{\beta}(\lambda)] &= \mathrm{MSE}[H(\lambda)y] \\
&= \mathrm{E}[\|H(\lambda)y - X\beta\|^2] \\
&= \mathrm{E}[\|H(\lambda)y - \mathrm{E}[H(\lambda)y] + \mathrm{E}[H(\lambda)y] - X\beta\|^2] \\
&= \mathrm{E}[\|H(\lambda)y - H(\lambda)\mathrm{E}[y]\|^2] + \|H(\lambda)\mathrm{E}[y] - X\beta\|^2 \\
&= \mathrm{tr}\,(\mathrm{Cov}(H(\lambda)y)) + \mathrm{bias}^2(H(\lambda)y) \\
&= \mathrm{tr}\,(H(\lambda)\mathrm{Cov}(y)H(\lambda)^{\mathrm{T}}) + \|H(\lambda)X\beta - X\beta\|^2 \\
&= \mathrm{tr}\,(H(\lambda) \cdot \sigma^2 I_n \cdot H(\lambda)^{\mathrm{T}}) + \|H(\lambda)X\beta - X\beta\|^2 \\
&= \sigma^2\,\mathrm{tr}\{X(X^{\mathrm{T}}X + \lambda I)^{-1}X^{\mathrm{T}}X(X^{\mathrm{T}}X + \lambda I)^{-1}X^{\mathrm{T}}\} + \|X(X^{\mathrm{T}}X + \lambda I)^{-1}X^{\mathrm{T}}X\beta - X\beta\|^2 \\
&= \sigma^2\,\mathrm{tr}\{(X^{\mathrm{T}}X + \lambda I)^{-1}X^{\mathrm{T}}X(X^{\mathrm{T}}X + \lambda I)^{-1}X^{\mathrm{T}}X\} + \| - \lambda X(X^{\mathrm{T}}X + \lambda I)^{-1}\beta\|^2 \\
&= \sigma^2\,\mathrm{tr}\{(X^{\mathrm{T}}X + \lambda I)^{-1}X^{\mathrm{T}}X(X^{\mathrm{T}}X + \lambda I)^{-1}X^{\mathrm{T}}X\} + \lambda^2\beta^{\mathrm{T}}(X^{\mathrm{T}}X + \lambda I)^{-1}X^{\mathrm{T}}X(X^{\mathrm{T}}X + \lambda I)^{-1}\beta
\end{aligned}$$

设 $X^{\mathrm{T}}X \in \mathbb{R}^{(p+1)\times(p+1)}$ 的谱分解为 $X^{\mathrm{T}}X = UDU^{\mathrm{T}}$
其中 $U \in \mathbb{R}^{(p+1)\times(p+1)}$ 为正交阵，$D = \mathrm{diag}\{d_1, \ldots, d_{p+1}\}$ 为对角阵.
于是我们有:

$$\begin{aligned}
&\mathrm{MSE}[X\hat{\beta}(\lambda)] \\
&= \sigma^2\,\mathrm{tr}\{(X^{\mathrm{T}}X + \lambda I)^{-1}X^{\mathrm{T}}X(X^{\mathrm{T}}X + \lambda I)^{-1}X^{\mathrm{T}}X\} + \lambda^2\beta^{\mathrm{T}}(X^{\mathrm{T}}X + \lambda I)^{-1}X^{\mathrm{T}}X(X^{\mathrm{T}}X + \lambda I)^{-1}\beta \\
&= \sigma^2\,\mathrm{tr}\{(UDU^{\mathrm{T}} + \lambda I)^{-1}UDU^{\mathrm{T}}(UDU^{\mathrm{T}} + \lambda I)^{-1}UDU^{\mathrm{T}}\} + \lambda^2\beta^{\mathrm{T}}(UDU^{\mathrm{T}} + \lambda I)^{-1}UDU^{\mathrm{T}}(UDU^{\mathrm{T}} + \lambda I)^{-1}\beta \\
&= \sigma^2\,\mathrm{tr}\{U(D + \lambda I)^{-1}D(D + \lambda I)^{-1}DU^{\mathrm{T}}\} + \lambda^2\beta^{\mathrm{T}}U(D + \lambda I)^{-1}D(D + \lambda I)^{-1}U^{\mathrm{T}}\beta \\
&= \sigma^2\,\mathrm{tr}\,((D + \lambda I)^{-2}D^2) + \lambda^2\beta^{\mathrm{T}}U(D + \lambda I)^{-2}DU^{\mathrm{T}}\beta
\end{aligned}$$

注意到最小二乘估计量 $\hat{\beta}=(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y=\hat{\beta}(0)$，即 Ridge 估计量 $\lambda=0$ 的情形.
因此要证明存在某个 $\lambda>0$ 使得 $\mathrm{MSE}(X\hat{\beta}(\lambda))<\mathrm{MSE}(X\hat{\beta})$，
我们只要证明 $\mathrm{MSE}(X\hat{\beta}(\lambda))$ 关于 λ 的导数在 $\lambda=0$ 处为负值即可.

$$\begin{aligned}\frac{d}{d\lambda}\mathrm{MSE}(X\hat{\beta}(\lambda))&=\frac{d}{d\lambda}\left\{\sigma^2\,\mathrm{tr}\left((D+\lambda I)^{-2}D^2\right)+\lambda^2\beta^{\mathrm{T}}U(D+\lambda I)^{-2}DU^{\mathrm{T}}\beta\right\}\\&=-2\sigma^2\,\mathrm{tr}\left((D+\lambda I)^{-3}D^2\right)+2\lambda\beta^{\mathrm{T}}U(D+\lambda I)^{-2}DU^{\mathrm{T}}\beta+\lambda^2\cdot(-2\beta^{\mathrm{T}}U(D+\lambda I)^{-3}DU^{\mathrm{T}}\beta)\\\hline\frac{d}{d\lambda}\mathrm{MSE}(X\hat{\beta}(\lambda))\Big|_{\lambda=0}&=-2\sigma^2\,\mathrm{tr}\left(D^{-3}D^2\right)+0+0\\&=-2\sigma^2\,\mathrm{tr}\left(D^{-1}\right)\\&<0\end{aligned}$$

命题得证.

Problem 3

Consider an ridge estimator for $\beta\in\mathbb{R}^{p+1}$ in linear regression model:

$$\hat{\beta}(\Lambda)=(X^{\mathrm{T}}X+U\Lambda U^{\mathrm{T}})^{-1}X^{\mathrm{T}}y$$

where $U\in\mathbb{R}^{(p+1)\times(p+1)}$ is an orthogonal matrix such that $U^{\mathrm{T}}X^{\mathrm{T}}XU=D=\mathrm{diag}\{d_1,\ldots,d_{p+1}\}$
and $\Lambda=\mathrm{diag}\{\lambda_1,\ldots,\lambda_{p+1}\}$ is a diagonal matrix.
Prove that there exists $\Lambda\succ 0$ such that $\mathrm{MSE}(\hat{\beta}(\Lambda))<\mathrm{MSE}(\hat{\beta})$ (where $\hat{\beta}=(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y$)

Proof:
考虑 $\hat{\beta}(\Lambda)=(X^{\mathrm{T}}X+U\Lambda U^{\mathrm{T}})^{-1}X^{\mathrm{T}}y$ 的均方误差:

$$\begin{aligned}&\mathrm{MSE}[\hat{\beta}(\Lambda)]\\&=\mathrm{E}[\|\hat{\beta}(\Lambda)-\beta\|^2]\\&=\mathrm{E}[\|\hat{\beta}(\Lambda)-\mathrm{E}[\hat{\beta}(\Lambda)]+\mathrm{E}[\hat{\beta}(\Lambda)]-\beta\|^2]\\&=\mathrm{E}[\|\hat{\beta}(\Lambda)-\mathrm{E}[\hat{\beta}(\Lambda)]\|^2]+\|\mathrm{E}[\hat{\beta}(\Lambda)]-\beta\|^2\\&=\mathrm{tr}\left(\mathrm{Cov}(\hat{\beta}(\Lambda))\right)+\mathrm{bias}^2(\hat{\beta}(\Lambda))\quad(\text{note that }\hat{\beta}(\Lambda)=(X^{\mathrm{T}}X+U\Lambda U^{\mathrm{T}})^{-1}X^{\mathrm{T}}y)\\&=\mathrm{tr}\{(X^{\mathrm{T}}X+U\Lambda U^{\mathrm{T}})^{-1}X^{\mathrm{T}}\cdot\mathrm{Cov}(y)\cdot[(X^{\mathrm{T}}X+U\Lambda U^{\mathrm{T}})^{-1}X^{\mathrm{T}}]^{\mathrm{T}}\}+\|(X^{\mathrm{T}}X+U\Lambda U^{\mathrm{T}})^{-1}X^{\mathrm{T}}\mathrm{E}[y]-\beta\|^2\\&=\mathrm{tr}\{(X^{\mathrm{T}}X+U\Lambda U^{\mathrm{T}})^{-1}X^{\mathrm{T}}\cdot\sigma^2I_n\cdot[(X^{\mathrm{T}}X+U\Lambda U^{\mathrm{T}})^{-1}X^{\mathrm{T}}]^{\mathrm{T}}\}+\|(X^{\mathrm{T}}X+U\Lambda U^{\mathrm{T}})^{-1}X^{\mathrm{T}}X\beta-\beta\|^2\\&=\sigma^2\,\mathrm{tr}\{(X^{\mathrm{T}}X+U\Lambda U^{\mathrm{T}})^{-1}X^{\mathrm{T}}X(X^{\mathrm{T}}X+U\Lambda U^{\mathrm{T}})^{-1}\}+\|-(X^{\mathrm{T}}X+U\Lambda U^{\mathrm{T}})^{-1}U\Lambda U^{\mathrm{T}}\beta\|^2\\&=\sigma^2\,\mathrm{tr}\{(X^{\mathrm{T}}X+U\Lambda U^{\mathrm{T}})^{-2}X^{\mathrm{T}}X\}+\beta^{\mathrm{T}}U\Lambda U^{\mathrm{T}}(X^{\mathrm{T}}X+U\Lambda U^{\mathrm{T}})^{-2}U\Lambda U^{\mathrm{T}}\beta\quad(\text{note that }X^{\mathrm{T}}X=UDU^{\mathrm{T}})\\&=\sigma^2\,\mathrm{tr}\{(UDU^{\mathrm{T}}+U\Lambda U^{\mathrm{T}})^{-2}UDU^{\mathrm{T}}\}+\beta^{\mathrm{T}}U\Lambda U^{\mathrm{T}}(UDU^{\mathrm{T}}+U\Lambda U^{\mathrm{T}})^{-2}U\Lambda U^{\mathrm{T}}\beta\\&=\sigma^2\,\mathrm{tr}\{U(D+\Lambda)^{-2}DU^{\mathrm{T}}\}+\beta^{\mathrm{T}}U\Lambda(D+\Lambda)^{-2}\Lambda U^{\mathrm{T}}\beta\\&=\sigma^2\,\mathrm{tr}\{(D+\Lambda)^{-2}D\}+\beta^{\mathrm{T}}U(D+\Lambda)^{-2}\Lambda^2U^{\mathrm{T}}\beta\end{aligned}$$

注意到最小二乘估计量 $\hat{\beta}=(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y=\hat{\beta}(0_{(p+1)\times(p+1)})$，即 Ridge 估计量 $\Lambda=0_{(p+1)\times(p+1)}$ 的情形.
因此要证明存在某个 $\Lambda\succ 0$ 使得 $\mathrm{MSE}(\hat{\beta}(\Lambda))<\mathrm{MSE}(\hat{\beta})$，
我们只要证明 $\mathrm{MSE}(\hat{\beta}(\Lambda))$ 关于 Λ 的梯度在 $\Lambda=0_{(p+1)\times(p+1)}$ 处为负定矩阵即可.

$$\begin{aligned}\nabla_\Lambda\mathrm{MSE}(\hat{\beta}(\Lambda))&=\nabla_\Lambda\{\sigma^2\,\mathrm{tr}\left((D+\Lambda)^{-2}D\right)+\beta^{\mathrm{T}}U(D+\Lambda)^{-2}\Lambda^2U^{\mathrm{T}}\beta\}\\&=-2\sigma^2(D+\Lambda)^{-3}D+[(D+\Lambda)^{-2}\cdot 2\Lambda+(-2(D+\Lambda)^{-3})\Lambda^2]\cdot\mathrm{diag}\{(U^{\mathrm{T}}\beta)\odot(U^{\mathrm{T}}\beta)\}\\&=-2\sigma^2(D+\Lambda)^{-3}D+2(D+\Lambda)^{-2}\Lambda[I-(D+\Lambda)^{-1}\Lambda]\cdot\mathrm{diag}\{(U^{\mathrm{T}}\beta)\odot(U^{\mathrm{T}}\beta)\}\\\hline\nabla_\Lambda\mathrm{MSE}(\hat{\beta}(\Lambda))\Big|_{\Lambda=0_{(p+1)\times(p+1)}}&=-2\sigma^2D^{-3}D+0_{(p+1)\times(p+1)}\\&=-2\sigma^2D^{-2}\\&\prec 0\end{aligned}$$

命题得证.

Problem 4

Suppose that only two values, $x = 0, 1$ are observed.
For $x = 0$, there are 10 successes in 10 trials.
For $x = 1$, there are 5 successes in 10 trials.
Show that the logistic regression MLEs $\hat{\alpha}$, $\hat{\beta}$ does not exist.

Solution:
更改记号: 将 α 改为 β_0，将 β 改为 β_1，记 $\beta = [\beta_0, \beta_1]^{\mathrm{T}}$
记样本为:

$$(x_i, y_i) = \begin{cases} (0,1) & \text{if } i = 1, \dots, 10 \\ (1,1) & \text{if } i = 11, \dots, 15 \\ (1,0) & \text{if } i = 16, \dots, 20 \end{cases}$$

因此设计矩阵和响应变量观测可以表示为:

$$X = \begin{bmatrix} 1_{10} & 0_{10} \\ 1_{10} & 1_{10} \end{bmatrix} \quad y = \begin{bmatrix} 1_{15} \\ 0_5 \end{bmatrix}$$

记 $n = 20$
假设 $Y_i \ (i = 1, \dots, n)$ 服从以下 Bernoulli 分布:

$$Y_i \sim \mathrm{B}(1, p_i) \text{ i.e. } Y_i = \begin{cases} 1, & p_i \\ 0, & 1 - p_i \end{cases}$$
$$\text{where } p_i := \mathrm{P}\{Y = 1 | x_i\} = \sigma(\beta^{\mathrm{T}} x_i)$$

因此 $Y = [Y_1, \dots, Y_n]^{\mathrm{T}}$ 的联合概率密度函数为:

$$\begin{aligned} f(Y) &:= \prod_{i=1}^{n} p_i^{Y_i} (1 - p_i)^{1 - Y_i} \\ &= \prod_{i=1}^{n} (\sigma(\beta^{\mathrm{T}} x_i))^{Y_i} (1 - \sigma(\beta^{\mathrm{T}} x_i))^{1 - Y_i} \end{aligned}$$

对数似然函数为:

$$\begin{aligned} \mathcal{L}(y | X, \beta) &= \log\left(\prod_{i=1}^{n} p_i^{y_i} (1 - p_i)^{1 - y_i} \right) \\ &= \sum_{i=1}^{n} \{ y_i \log(p_i) + (1 - y_i) \log(1 - p_i) \} \\ &= \sum_{i=1}^{n} \left\{ y_i \log\left(\sigma(\beta^{\mathrm{T}} x_i)\right) + (1 - y_i) \log\left(1 - \sigma(\beta^{\mathrm{T}} x_i)\right) \right\} \\ &= \sum_{i=1}^{n} \left\{ y_i \log\left(\sigma(\beta^{\mathrm{T}} x_i)\right) + (1 - y_i) \log\left(\sigma(-\beta^{\mathrm{T}} x_i)\right) \right\} \\ &= \sum_{i=1}^{n} \left\{ y_i \log\left(\frac{1}{1 + \exp(-\beta^{\mathrm{T}} x_i)} \right) - y_i \log\left(\frac{1}{1 + \exp(\beta^{\mathrm{T}} x_i)} \right) + \log\left(\sigma(-\beta^{\mathrm{T}} x_i)\right) \right\} \\ &= \sum_{i=1}^{n} \left\{ y_i \log\left(\frac{1 + \exp(\beta^{\mathrm{T}} x_i)}{1 + \exp(-\beta^{\mathrm{T}} x_i)} \right) + \log\left(\sigma(-\beta^{\mathrm{T}} x_i)\right) \right\} \\ &= \sum_{i=1}^{n} \left\{ y_i \log\left(\exp\left(\beta^{\mathrm{T}} x_i\right)\right) + \log\left(\sigma(-\beta^{\mathrm{T}} x_i)\right)) \right\} \\ &= \sum_{i=1}^{n} \left\{ y_i (\beta^{\mathrm{T}} x_i) + \log\left(\sigma(-\beta^{\mathrm{T}} x_i)\right) \right\} \\ &= y^{\mathrm{T}} X \beta + 1_n^{\mathrm{T}} \log(\sigma(-X\beta)) \end{aligned}$$

其关于 β 的梯度为:

$$
\begin{aligned}
\nabla_\beta \mathcal{L}(y|X,\beta) &= \nabla_\beta \{y^{\mathrm{T}} X\beta + 1_n^{\mathrm{T}} \log(\sigma(-X\beta))\} \\
&= X^{\mathrm{T}} y + \sum_{i=1}^n \nabla_\beta \log\left(\sigma(-\beta^{\mathrm{T}} x_i)\right) \\
&= X^{\mathrm{T}} y + \sum_{i=1}^n \frac{1}{\sigma(-\beta^{\mathrm{T}} x_i)} \sigma(-\beta^{\mathrm{T}} x_i)[1-\sigma(-\beta^{\mathrm{T}} x_i)] \cdot (-x_i) \\
&= X^{\mathrm{T}} y - \sum_{i=1}^n \sigma(\beta^{\mathrm{T}} x_i) x_i \\
&= X^{\mathrm{T}} y - X^{\mathrm{T}} \sigma(X\beta) \\
&= X^{\mathrm{T}} (y - \sigma(X\beta))
\end{aligned}
$$

要证明在给定样本下 Logistic 回归没有极大似然解，
等价于证明优化问题 $\max_{\beta\in\mathbb{R}^2} \mathcal{L}(y|X,\beta)$ 无解，
只需证明 $\mathcal{L}(y|X,\beta)$ 在 $\mathbb{R}^2$ 中没有驻点，
即要证明 $\nabla_\beta \mathcal{L}(y|X,\beta) = X^{\mathrm{T}}(y-\sigma(X\beta))$ 在 $\mathbb{R}^2$ 中没有零点，
即要证明对于任意 $\beta \in \mathbb{R}^2$，$y - \sigma(X\beta)$ 都不与 X^{T} 的列空间正交.

回忆起设计矩阵和响应变量观测为:

$$
(x_i, y_i) = \begin{cases} (0,1) & \text{if } i = 1,\dots,10 \\ (1,1) & \text{if } i = 11,\dots,15 \\ (1,0) & \text{if } i = 16,\dots,20 \end{cases}
$$

$$
X = \begin{bmatrix} 1_{10} & 0_{10} \\ 1_{10} & 1_{10} \end{bmatrix} \quad y = \begin{bmatrix} 1_{15} \\ 0_5 \end{bmatrix}
$$

我们记:

$$
\begin{aligned}
a &:= \sigma\left(\begin{bmatrix} 1 \\ 0 \end{bmatrix}^{\mathrm{T}} \beta\right) \\
b &:= \sigma\left(\begin{bmatrix} 1 \\ 1 \end{bmatrix}^{\mathrm{T}} \beta\right)
\end{aligned}
$$

根据 Logistic 函数 $\sigma(\cdot)$ 的性质可知 $a, b \in (0,1)$
于是我们有:

$$
y_i - \sigma(x_i^{\mathrm{T}} \beta) = \begin{cases} 1-a & \text{if } i = 1,\dots,10 \\ 1-b & \text{if } i = 11,\dots,15 \\ -b & \text{if } i = 16,\dots,20 \end{cases}
$$

因此方程 $X^{\mathrm{T}}(y-\sigma(X\beta)) = 0_2$ 可以表示为:

$$
\begin{cases} 10(1-a) + 5(1-b) + 5\cdot(-b) = 15 - 10a - 10b = 0 \\ 5(1-b) + 5\cdot(-b) = 5 - 10b = 0 \end{cases}
$$

解得 $\begin{cases} a = 1 \\ b = \frac{1}{2} \end{cases}$
这与 $a, b \in (0,1)$ 的事实相矛盾，因此不存在 $\beta \in \mathbb{R}^2$ 使得 $X^{\mathrm{T}}(y-\sigma(X\beta)) = 0_2$
结合前面的推理可知在本题所给的样本下，Logistic 回归不存在极大似然解.

Problem 5

Consider the maximation of the ridge penalized loglikelihood of logistic regression:

$$
\begin{aligned}
\mathcal{L}(y|X,\beta,\lambda) &= y^{\mathrm{T}} X\beta + 1_n^{\mathrm{T}} \log(\sigma(-X\beta)) - \frac{1}{2}\lambda\|\beta\|^2 \\
&= \sum_{i=1}^n \left\{ y_i(\beta^{\mathrm{T}} x_i) + \log\left(\sigma(-\beta^{\mathrm{T}} x_i)\right)\right\} - \frac{1}{2}\lambda\|\beta\|^2 \\
&= \sum_{i=1}^n \left\{ y_i(\beta^{\mathrm{T}} x_i) - \log\left(1 + \exp\left(\beta^{\mathrm{T}} x_i\right)\right)\right\} - \frac{1}{2}\lambda\|\beta\|^2
\end{aligned}
$$

where $X = [x_1, \ldots, x_n]^{\mathrm{T}} \in \mathbb{R}^{n\times(p+1)}$, $\beta \in \mathbb{R}^{p+1}$ and $\lambda > 0$
Derive the Newton algorithm.
(Hint: it is similar to the iteratively re-weighted least squares algorithm for unpenalized logistic regression)

Solution:
首先求解 $\mathcal{L}(y|X,\beta,\lambda)$ 关于 β 的梯度:

$$\begin{aligned}
\nabla_\beta \mathcal{L}(y|X,\beta,\lambda) &= \nabla_\beta\{y^{\mathrm{T}}X\beta + 1_n^{\mathrm{T}}\log\left(\sigma(-X\beta)\right)\} - \lambda\beta \\
&= X^{\mathrm{T}}y + \sum_{i=1}^{n}\nabla_\beta \log\left(\sigma(-\beta^{\mathrm{T}}x_i)\right) - \lambda\beta \\
&= X^{\mathrm{T}}y + \sum_{i=1}^{n}\frac{1}{\sigma(-\beta^{\mathrm{T}}x_i)}\sigma(-\beta^{\mathrm{T}}x_i)[1-\sigma(-\beta^{\mathrm{T}}x_i)]\cdot(-x_i) - \lambda\beta \\
&= X^{\mathrm{T}}y - \sum_{i=1}^{n}\sigma(\beta^{\mathrm{T}}x_i)x_i - \lambda\beta \\
&= X^{\mathrm{T}}y - X^{\mathrm{T}}\sigma(X\beta) - \lambda\beta \\
&= X^{\mathrm{T}}(y - \sigma(X\beta)) - \lambda\beta
\end{aligned}$$

其次求解 $\mathcal{L}(y|X,\beta,\lambda)$ 关于 β 的 Hesse 矩阵:

$$\begin{aligned}
\nabla_\beta^2 \mathcal{L}(y|X,\beta,\lambda) &= \frac{\partial}{\partial\beta}\{\nabla_\beta\mathcal{L}(y|X,\beta,\lambda)\} \\
&= \frac{\partial}{\partial\beta}\{X^{\mathrm{T}}(y-\sigma(X\beta)) - \lambda\beta\} \\
&= -X^{\mathrm{T}}\cdot\mathrm{diag}\{\sigma(X\beta)\odot(1_n - \sigma(X\beta))\}\cdot X - \lambda I_{p+1}
\end{aligned}$$

这样我们就可以给出加入 l_2 惩罚项的纯 Newton 法的迭代算法:

- ① 初始化 $\beta^{(0)} = 0_{d+1}$
- ② 然后迭代更新参数直至达到某个预设定的停止条件:

$$\begin{aligned}
p^{(k)} &= \sigma(X\beta^{(k)}) \\
\nabla_\beta\mathcal{L}(y|X,\beta,\lambda) &= X^{\mathrm{T}}(y - p^{(k)}) - \lambda\beta^{(k)} \\
\nabla_\beta^2\mathcal{L}(y|X,\beta,\lambda) &= -X^{\mathrm{T}}\cdot\mathrm{diag}\{(1_n - p^{(k)})\odot p^{(k)}\}\cdot X - \lambda I_{p+1} \\
d^{(k)} &= (\nabla_\beta^2\mathcal{L}(y|X,\beta,\lambda))^{-1}\nabla_\beta\mathcal{L}(y|X,\beta,\lambda) \\
\beta^{(k+1)} &= \beta^{(k)} + d^{(k)}
\end{aligned}$$

(注意这是一个最大化问题，因此 $d^{(k)}$ 这里代表 "上升方向"，对应最小化问题中的下降方向)

写成加权最小二乘格式即为:

$$\begin{aligned}
\beta^{(k+1)} &= \beta^{(k)} + d^{(k)} \\
&= \beta^{(k)} + (\nabla_\beta^2\mathcal{L}(y|X,\beta,\lambda))^{-1}\nabla_\beta\mathcal{L}(y|X,\beta,\lambda) \\
&= \beta^{(k)} + [-X^{\mathrm{T}}\cdot\mathrm{diag}\{(1_n - p^{(k)})\odot p^{(k)}\}\cdot X - \lambda I_{p+1}]^{-1}[X^{\mathrm{T}}(y-p^{(k)}) - \lambda\beta^{(k)}] \\
&= \beta^{(k)} + (X^{\mathrm{T}}W_kX + \lambda I_{p+1})^{-1}X^{\mathrm{T}}W_kz^{(k)}
\end{aligned}$$

其中 $W_k = \mathrm{diag}\{(1_n - p^{(k)})\odot p^{(k)}\} = \mathrm{diag}\{(1_n - \sigma(X\beta^{(k)}))\odot\sigma(X\beta^{(k)})\}$
而 $z^{(k)} = W_k^{-1}[(y - p^{(k)}) - X(X^{\mathrm{T}}X)^{-1}\beta^{(k)}]$

The End