Q1

因为特征排列顺序并无特殊限制，不妨取 x_p 为因变量，其它特征为自变量，建立多元线性回归模型

$$\boldsymbol{x}_p = \alpha_1\boldsymbol{x}_1 + \alpha_2\boldsymbol{x}_2 + \cdots + \alpha_{p-1}\boldsymbol{x}_{p-1} + \boldsymbol{\epsilon}$$

令 $\boldsymbol{X}_t = (\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_{p-1})$，$\boldsymbol{X}_t$ 是该回归模型的构造矩阵

$$\Rightarrow \hat{\boldsymbol{x}}_p = \boldsymbol{X}_t\hat{\boldsymbol{\beta}}_t = \boldsymbol{X}_t(\boldsymbol{X}_t'\boldsymbol{X}_t)^{-1}\boldsymbol{X}_t'\boldsymbol{x}_p$$

记 $\boldsymbol{H} = \boldsymbol{X}_t(\boldsymbol{X}_t'\boldsymbol{X}_t)^{-1}\boldsymbol{X}_t'$，$\boldsymbol{H}$ 是一个帽子矩阵

回归系数 $R_p^2 = \dfrac{SS_{Rp}}{SS_{Tp}}$

其中

$$SS_{Rp} = \sum_{i=1}^{n}(\hat{x}_{ip} - \overline{x}_p)^2 = \sum_{i=1}^{n}\hat{x}_{ip}^2 = \hat{\boldsymbol{x}}_p'\hat{\boldsymbol{x}}_p = (\boldsymbol{H}\boldsymbol{x}_p)'(\boldsymbol{H}\boldsymbol{x}_p) = \boldsymbol{x}_p'\boldsymbol{H}'\boldsymbol{H}\boldsymbol{x}_p = \boldsymbol{x}_p'\boldsymbol{H}\boldsymbol{x}_p$$

$$SS_{Tp} = \sum_{i=1}^{n}(x_{ip} - \overline{x}_p)^2 = \sum_{i=1}^{n}x_{ip}^2 = 1$$

$$\Rightarrow R_p^2 = \boldsymbol{x}_p'\boldsymbol{H}\boldsymbol{x}_p$$

因为 $\boldsymbol{X}_s = (\boldsymbol{X}_t, \boldsymbol{x}_p)$

$$\boldsymbol{X}_s'\boldsymbol{X}_s = \begin{pmatrix}\boldsymbol{X}_t' \\ \boldsymbol{x}_p'\end{pmatrix}\begin{pmatrix}\boldsymbol{X}_t & \boldsymbol{x}_p\end{pmatrix} = \begin{pmatrix}\boldsymbol{X}_t'\boldsymbol{X}_t & \boldsymbol{X}_t'\boldsymbol{x}_p \\ \boldsymbol{x}_p'\boldsymbol{X}_t & \boldsymbol{x}_p'\boldsymbol{x}_p\end{pmatrix}$$

记 $\boldsymbol{X}_t'\boldsymbol{X}_t = \boldsymbol{A}_{11}$，$\boldsymbol{X}_t'\boldsymbol{x}_p = \boldsymbol{A}_{12}$，$\boldsymbol{x}_p'\boldsymbol{X}_t = \boldsymbol{A}_{21}$，$\boldsymbol{x}_p'\boldsymbol{x}_p = \boldsymbol{A}_{22}$

$$\Rightarrow (\boldsymbol{X}_s'\boldsymbol{X}_s)^{-1} = \begin{pmatrix}\boldsymbol{A}_{11}^{-1} + \boldsymbol{A}_{11}^{-1}\boldsymbol{A}_{12}\boldsymbol{M}^{-1}\boldsymbol{A}_{21}\boldsymbol{A}_{11}^{-1} & -\boldsymbol{A}_{11}^{-1}\boldsymbol{A}_{12}\boldsymbol{M}^{-1} \\ -\boldsymbol{M}^{-1}\boldsymbol{A}_{21}\boldsymbol{A}_{11}^{-1} & \boldsymbol{M}^{-1}\end{pmatrix}$$

其中 $\boldsymbol{M} = \boldsymbol{A}_{22} - \boldsymbol{A}_{21}\boldsymbol{A}_{11}^{-1}\boldsymbol{A}_{12}$

因为我们关心的是 VIF_p，所以只需求解 $\boldsymbol{M}^{-1}$

$$\boldsymbol{M} = \boldsymbol{A}_{22} - \boldsymbol{A}_{21}\boldsymbol{A}_{11}^{-1}\boldsymbol{A}_{12} = \boldsymbol{x}_p'\boldsymbol{x}_p - \boldsymbol{x}_p'\boldsymbol{X}_t(\boldsymbol{X}_t'\boldsymbol{X}_t)^{-1}\boldsymbol{X}_t'\boldsymbol{x}_p = \boldsymbol{x}_p'\boldsymbol{x}_p - \boldsymbol{x}_p'\boldsymbol{H}\boldsymbol{x}_p = 1 - R_p^2$$

所以 $\boldsymbol{M}^{-1} = (1 - R_p^2)^{-1}$

因为 VIF_p 为 $(\boldsymbol{X}_s'\boldsymbol{X}_s)^{-1}$ 的第 p 个对角线元素

$$\Rightarrow VIF_p = \boldsymbol{M}^{-1} = \frac{1}{1 - R_p^2}$$

因为特征顺序对模型回归结果无影响，因此将任一特征作为第 p 个特征对结果无影响。

即该结论具有一般性，可推广为

$$VIF_j = \frac{1}{1-R_j^2}$$

Q2

由最小二乘的性质可得

$$E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}, \quad \mathrm{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$$

$$\begin{aligned} MSE(\hat{\boldsymbol{\beta}}) &= E(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})'(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}) \\ &= E(\hat{\boldsymbol{\beta}}-E\hat{\boldsymbol{\beta}}+E\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})'(\hat{\boldsymbol{\beta}}-E\hat{\boldsymbol{\beta}}+E\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}) \\ &= E(\hat{\boldsymbol{\beta}}-E\hat{\boldsymbol{\beta}})'(\hat{\boldsymbol{\beta}}-E\hat{\boldsymbol{\beta}}) \\ &= \mathrm{tr}\Big(\mathrm{Cov}(\hat{\boldsymbol{\beta}})\Big) \\ &= \mathrm{tr}\Big(\mathrm{Var}(\hat{\boldsymbol{\beta}})\Big) \\ &= \mathrm{tr}\big(\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}\big) \\ &= \sigma^2\,\mathrm{tr}\,\big(\boldsymbol{X}'\boldsymbol{X}\big)^{-1} \end{aligned}$$

设 $\boldsymbol{X}'\boldsymbol{X}$ 的特征值为 $\lambda_1, \lambda_2, \cdots, \lambda_{p+1}$

根据矩阵的性质，有 $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ 的特征值为 $\lambda_1^{-1}, \lambda_2^{-1}, \cdots, \lambda_{p+1}^{-1}$，因此

$$\mathrm{tr}\,\big(\boldsymbol{X}'\boldsymbol{X}\big)^{-1} = \sum_{i=1}^{p+1}\frac{1}{\lambda_i} \quad \Rightarrow \quad MSE(\hat{\boldsymbol{\beta}}) = \sigma^2\sum_{i=1}^{p+1}\frac{1}{\lambda_i}$$

Q3

赤池信息量准则 AIC 定义为

$$\mathrm{AIC} = -2\ln(\text{模型最大似然}) + 2(\text{模型独立参数个数})$$

在线性回归模型中，假定数据满足以下分布

$$y \sim \mathcal{N}_n(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I})$$

$\boldsymbol{y} = (y_1, y_2, \cdots, y_n)'$ 的联合密度函数为

$$f(\boldsymbol{y}; \boldsymbol{\beta}, \sigma^2) = \frac{1}{(2\pi)^{n/2}|\sigma^2\boldsymbol{I}|^{1/2}}\exp\left(-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})'(\sigma^2\boldsymbol{I})^{-1}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})\right)$$

参数 $(\boldsymbol{\beta}, \sigma^2)$ 的似然函数为

$$L(\boldsymbol{\beta}, \sigma^2) = (2\pi)^{-n/2}(\sigma^2)^{-n/2}\exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})\right)$$

对数似然函数为

$$\ln L(\boldsymbol{\beta},\sigma^2) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2(\boldsymbol{y}-\boldsymbol{X\beta})'(\boldsymbol{y}-\boldsymbol{X\beta})}$$

- $\boldsymbol{\beta}$ 的最大似然估计

$$\frac{\partial \ln L(\boldsymbol{\beta},\sigma^2)}{\partial \boldsymbol{\beta}} = \frac{\partial(\frac{1}{2\sigma^2}(\boldsymbol{y}-\boldsymbol{X\beta})'(\boldsymbol{y}-\boldsymbol{X\beta}))}{\partial \boldsymbol{\beta}} = \frac{1}{\sigma^2}(-\boldsymbol{X})'(\boldsymbol{y}-\boldsymbol{X\beta}) = 0$$

$$\hat{\boldsymbol{\beta}}_{\mathrm{ML}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$$

- σ^2 的最大似然估计

$$\frac{\partial \ln L(\boldsymbol{\beta},\sigma^2)}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}(\boldsymbol{y}-\boldsymbol{X\beta})'(\boldsymbol{y}-\boldsymbol{X\beta}) = 0$$

$$\hat{\sigma}^2_{\mathrm{ML}} = \frac{1}{n}(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}_{\mathrm{ML}})'(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}_{\mathrm{ML}}) = \frac{1}{n}(\boldsymbol{y}-\hat{\boldsymbol{y}})'(\boldsymbol{y}-\hat{\boldsymbol{y}}) = \frac{1}{n}\boldsymbol{e}'\boldsymbol{e}$$

将参数估计代入，得到最大对数似然函数

$$\ln L(\boldsymbol{\beta},\sigma^2) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\frac{\boldsymbol{e}'\boldsymbol{e}}{n} - \frac{n}{2\boldsymbol{e}'\boldsymbol{e}}(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})'(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})$$

因此

$$\begin{aligned}
\mathrm{AIC} &= -2\ln L(\boldsymbol{\beta},\sigma^2) + 2(p+2) \\
&= n\ln(2\pi) + n\ln\frac{\boldsymbol{e}'\boldsymbol{e}}{n} + \frac{n}{\boldsymbol{e}'\boldsymbol{e}}(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})'(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}) + 2(p+2) \\
&= n\ln(2\pi) + n\ln\frac{\boldsymbol{e}'\boldsymbol{e}}{n} + \frac{n}{\boldsymbol{e}'\boldsymbol{e}}(\boldsymbol{y}-\hat{\boldsymbol{y}})'(\boldsymbol{y}-\hat{\boldsymbol{y}}) + 2(p+2) \\
&= n\ln(2\pi) + n\ln\left(\frac{SS_E}{n}\right) + n + 2(p+2)
\end{aligned}$$

Q4

已知岭回归估计是最小化带有 L_2 正则项的离差平方和的解，即

$$\hat{\boldsymbol{\beta}}(k) = \arg\min_{\boldsymbol{\beta}}(\boldsymbol{y}-\boldsymbol{X\beta})'(\boldsymbol{y}-\boldsymbol{X\beta}) + \lambda\boldsymbol{\beta}'\boldsymbol{\beta} \qquad \cdots\cdots(1)$$

以下通过贝叶斯统计中的最大后验估计证明上式

首先考虑线性回归模型

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$$

其中，$\boldsymbol{y}$ 是中心化后的响应变量，$\boldsymbol{X}$ 是标准化后的特征矩阵，$\boldsymbol{\beta}$ 是待估参数，误差项 $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0},\sigma^2\boldsymbol{I})$

在这个模型下，给定参数 $\boldsymbol{\beta}$，响应变量 $\boldsymbol{y}$ 的似然函数为

$$
\begin{aligned}
P(\boldsymbol{y}|\boldsymbol{\beta}) &= \frac{1}{(2\pi)^{n/2}|\sigma^2\boldsymbol{I}|^{1/2}}\exp\left(-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{X\beta})'(\sigma^2\boldsymbol{I})^{-1}(\boldsymbol{y}-\boldsymbol{X\beta})\right)\\
&= (2\pi)^{-n/2}(\sigma^2)^{-n/2}\exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{y}-\boldsymbol{X\beta})'(\boldsymbol{y}-\boldsymbol{X\beta})\right)
\end{aligned}
$$

对于 β_i，我们选择正态分布作为其先验分布

$$
\beta_i \sim \mathcal{N}(0,\tau^2), \quad i=1,\cdots,p
$$

即 $\boldsymbol{\beta}\sim\mathcal{N}(\boldsymbol{0},\tau^2\boldsymbol{I})$，所以有

$$
P(\boldsymbol{\beta}) = (2\pi)^{-p/2}(\tau^2)^{-p/2}\exp\left(-\frac{1}{2\tau^2}\boldsymbol{\beta}'\boldsymbol{\beta}\right)
$$

根据贝叶斯定理，后验分布为

$$
\begin{aligned}
P(\boldsymbol{\beta}|\boldsymbol{y}) &= P(\boldsymbol{y}|\boldsymbol{\beta})\cdot P(\boldsymbol{\beta})\\
&= (2\pi)^{-n/2}(\sigma^2)^{-n/2}\exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{y}-\boldsymbol{X\beta})'(\boldsymbol{y}-\boldsymbol{X\beta})\right)\cdot(2\pi)^{-p/2}(\tau^2)^{-p/2}\exp\left(-\frac{1}{2\tau^2}\boldsymbol{\beta}'\boldsymbol{\beta}\right)\\
&\propto \exp\left(-\frac{(\boldsymbol{y}-\boldsymbol{X\beta})'(\boldsymbol{y}-\boldsymbol{X\beta})}{2\sigma^2}-\frac{\boldsymbol{\beta}'\boldsymbol{\beta}}{2\tau^2}\right)
\end{aligned}
$$

要最大化后验分布，考虑最大化其对数函数

$$
L(\boldsymbol{\beta}) = -\frac{1}{2}\left(\frac{(\boldsymbol{y}-\boldsymbol{X\beta})'(\boldsymbol{y}-\boldsymbol{X\beta})}{\sigma^2}+\frac{\boldsymbol{\beta}'\boldsymbol{\beta}}{\tau^2}\right)
$$

可转化为最小化 $-2L(\boldsymbol{\beta})$

$$
\tilde{L}(\boldsymbol{\beta}) = -2L(\boldsymbol{\beta}) = \frac{1}{\sigma^2}(\boldsymbol{y}-\boldsymbol{X\beta})'(\boldsymbol{y}-\boldsymbol{X\beta})+\frac{1}{\tau^2}\boldsymbol{\beta}'\boldsymbol{\beta}
$$

令 $\dfrac{\sigma^2}{\tau^2}=\lambda$，所以得到后验分布的最大值

$$
\begin{aligned}
\arg\min_{\boldsymbol{\beta}}\tilde{L}(\boldsymbol{\beta}) &= \arg\min_{\boldsymbol{\beta}}\left(\frac{1}{\sigma^2}(\boldsymbol{y}-\boldsymbol{X\beta})'(\boldsymbol{y}-\boldsymbol{X\beta})+\frac{1}{\tau^2}\boldsymbol{\beta}'\boldsymbol{\beta}\right)\\
&= \arg\min_{\boldsymbol{\beta}}((\boldsymbol{y}-\boldsymbol{X\beta})'(\boldsymbol{y}-\boldsymbol{X\beta})+\lambda\boldsymbol{\beta}'\boldsymbol{\beta}) \quad \cdots\cdots(2)
\end{aligned}
$$

比较 (1), (2) 两式，形式相同，因此从贝叶斯统计的角度，岭回归可以看作是在进行线性回归时，通过对模型参数施加一个正态先验分布，从而引入了 L_2 正则化。