实验1　线性回归-岭回归

1. 实验目的

（1）掌握岭回归基本原理。

（2）应用岭回归模型解决问题。

2. 数据集介绍

在此次实验中,使用鲍鱼数据对鲍鱼的年龄进行预测，而鲍鱼的年龄是根据鲍鱼的环数来得到的。需要使用鲍鱼其他的特征（如性别、长度、直径、高度、整体重量、去壳重量、内脏重量和壳重量）来预测鲍鱼的年龄。数据集的各个数据维度特征如下：

特征名	类型	描述	单位
Sex	分类	性别（M雄性/F雌性/I幼体）	-
Length	连续	最长壳长	mm
Diameter	连续	垂直于长度的壳宽	mm
Height	连续	带内脏的壳高	mm
Whole_weight	连续	整体重量	g
Shucked_weight	连续	去壳后肉质重量	g
Viscera_weight	连续	内脏重量	g
Shell_weight	连续	干燥后壳重	g
Rings	整数	**目标变量：年龄标志**	个

3. 实验内容

编写程序实现岭回归模型，并解决现实应用问题。

4. 实验结果

4.1 岭回归的原理

岭回归是一种通过引入**L2正则化**（权重平方和）来解决多元线性回归中多重共线性问题的改进方法，它通过惩罚项 $\lambda\|w\|_2$ 约束模型参数，从而获得更稳定的解。与标准线性回归相比，岭回归能有效防止过拟合，尤其适用于特征相关性高的场景。

4.1.1 多元线性回归

核心思想：

- 用线性方程拟合多个自变量（特征）与因变量（目标）之间的关系：

  $$y = w_1x_1 + w_2x_2 + \cdots + w_dx_d + b$$

  其中 w 为权重（系数），b 为偏置（截距）

损失函数：

最小化**均方误差（MSE）**：

$$\ell(w) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

特点：

优点：

- 计算简单，解释性强。
- 解析解为 $w^* = (X^T X)^{-1} X^T y$（**当 $X^T X$ 可逆时才会有这个确定解**）。

缺点：

- 对**多重共线性**敏感（特征相关性高时，$X^T X$ 不可逆）。
- 容易**过拟合**（尤其是高维数据）。

4.1.2 Lasso的原理（L1正则化）

核心思想：

在损失函数中加入 **L1范数**（权重绝对值和）作为惩罚项，推动部分权重为0：

$$\ell(w) = \|y - Xw\|^2 + \lambda\|w\|_1$$

损失函数也是MSE。

特点：

优点：

- **自动特征选择**：稀疏解（无关特征的权重=0）。
- 适合**高维数据**（特征数 >> 样本数）。

缺点：

- 若 λ 太大，可能过度稀疏，丢失有用特征。
- 当特征高度相关时，可能随机选择一个而忽略其他。

4.1.2 岭回归（L2正则化）

核心思想：

在损失函数中加入 **L2范数**（权重平方和）作为惩罚项，限制权重幅度：

$$\ell(w) = \|y - Xw\|^2 + \lambda\|w\|_2^2$$

其中，$\|w\|_2^2$ 表示向量的二范数的平方，λ 是一个正则化参数，用于控制正则化项的重要性。损失函数也是MSE。岭回归的目标函数可表示为：

$$\min_w \|y - Xw\|_2^2 + \lambda\|w\|_2^2$$

特点：

优点：

- 解决**多重共线性**问题（强制 $X^T X + \lambda I$ 可逆）。
- 权重平滑，对异常值不敏感。

缺点：

- 不会产生稀疏解（所有特征都会被保留）。

4.1.4弹性网络（Elastic Net）

核心思想：

结合 **L1和L2正则化**，平衡稀疏性与稳定性：

$$\ell(w) = \|y - Xw\|^2 + \lambda_1 \|w\|_1 + \lambda_2 \|w\|_2^2$$

损失函数也是MSE。

特点：

优点：

- 同时具备Lasso的特征选择和Ridge的稳定性。
- 适合特征高度相关且需要稀疏性的场景。

缺点：

- 超参数（λ_1, λ_2）调优更复杂。

4.1.5 对比总结

方法	正则化类型	稀疏性	解决共线性	适用场景
多元线性回归	无	✗	✗	低维独立特征
Lasso回归	**L1范数**	✓	✗	**高维数据、特征选择**
岭回归	L2范数	✗	✓	共线性数据、防止过拟合
弹性网络	L1 + L2	✓	✓	高维共线性数据、需平衡稀疏与稳定

4.2 数据集分布分析

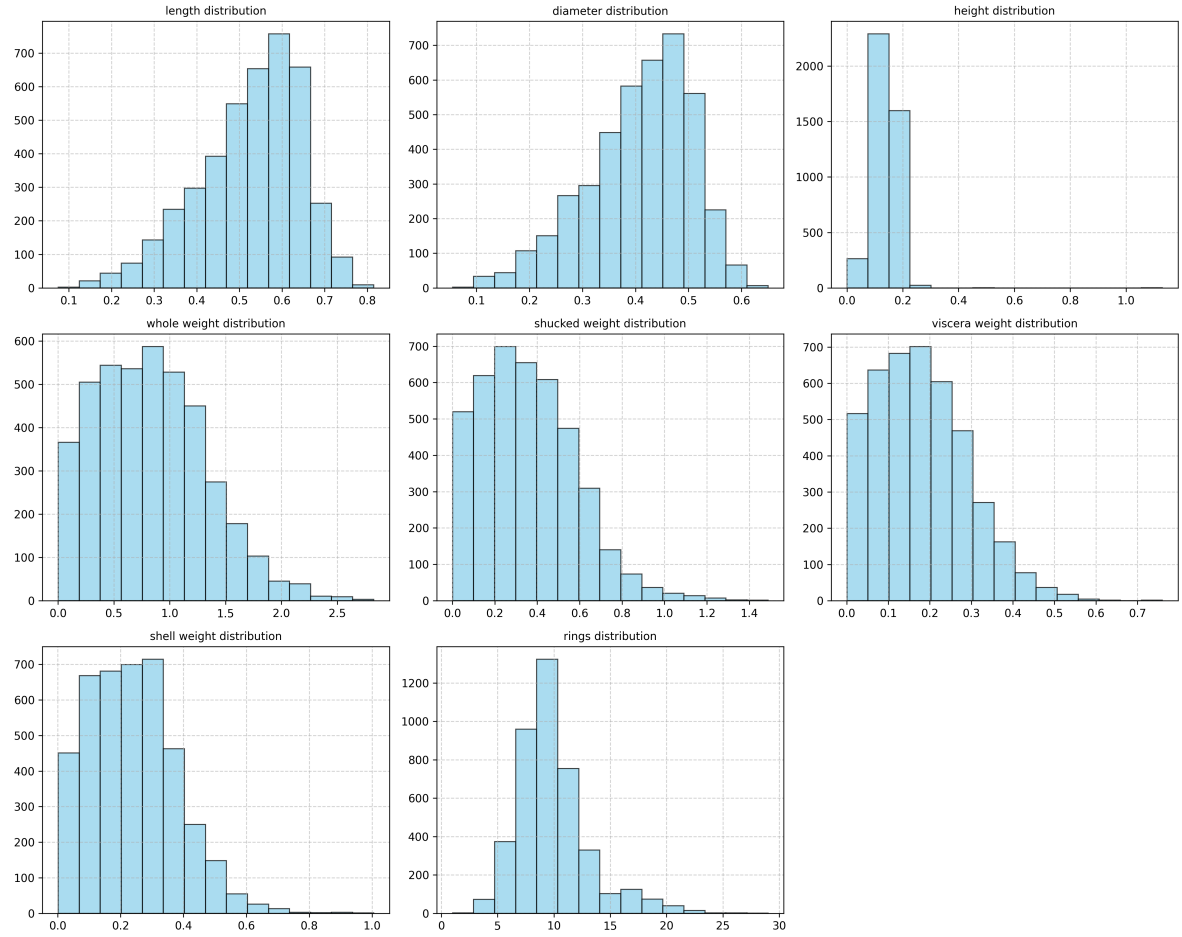

图1　数值特征分布直方图

图 1 展示了鲍鱼样本的形态学与重量指标特征分布规律：长度和直径呈对称钟形分布，高度则呈现左偏形态；重量特征中，整体重量为单峰右偏分布，去壳重量显示出独特的双峰结构，内脏重量分布相对集中，而壳重量呈现宽峰分布；目标变量环数的分布显著右偏，主要集中于环区间，同时存在少量高龄个体形成的长尾。各特征的分布范围和形态差异反映了鲍鱼生长过程中不同生理指标的变异规律。

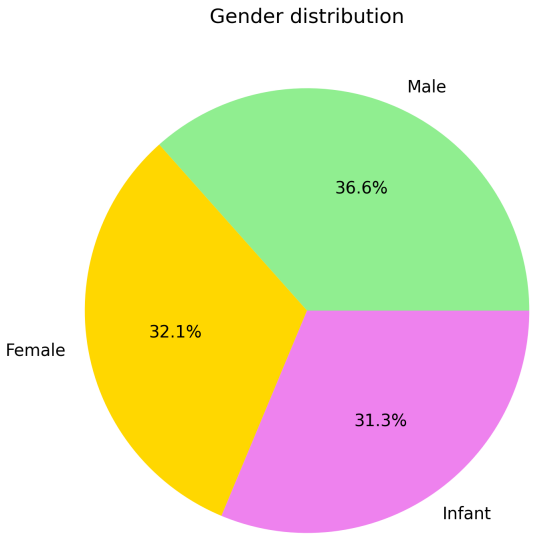

图2　性别分布饼图

图 2 为性别分布饼图直观呈现了鲍鱼样本的性别构成特征：三类性别群体（雄性36.6%、雌性32.1%、幼体31.3%）的比例分布相对均衡，最大差异不超过 5.3 个百分点。其中雄性占比略高，但三类群体的比例均落在30 %-37 %的相近区间，表明样本采集过程未出现明显的性别选择性偏差。

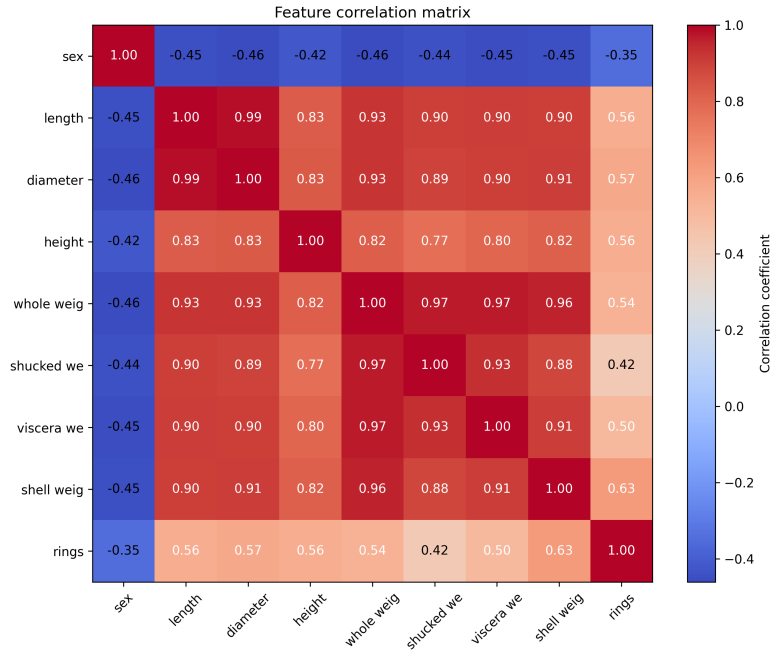

图3　特征相关性矩阵

图 3 为各个特征的相关性矩阵图。结果表明，形态学指标（长度、直径、高度）之间表现出极强的正相关性（R = 0.83-0.99），特别是长度与直径几乎完全共线性。同时，重量特征（整体重量、去壳重量、内脏重量、壳重量）与形态学特征保持高度正相关，其中整体重量与去壳重量的相关性最强。

4.3 建模结果

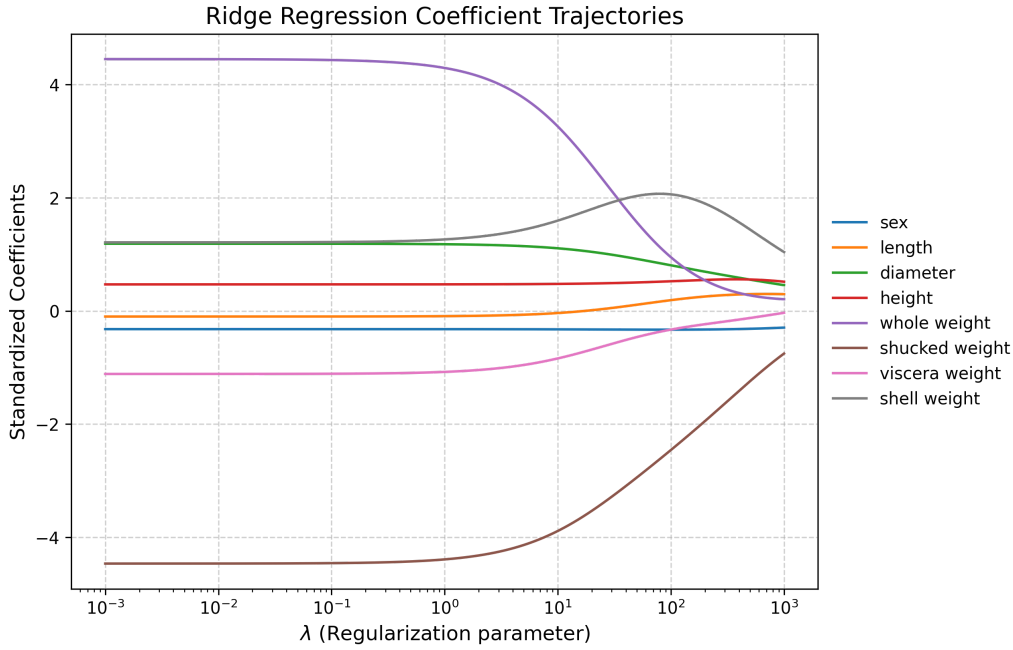

图4　特征系数随 λ 变化变化情况

图 4 为岭回归系数轨迹图，清晰地展示了各特征系数随正则化强度λ的变化规律：当λ值较小时（10^{-3}至10^{-1}区间），壳重表现出显著的正向影响，而去壳重量则呈现强烈的负向作用，表明在弱正则化条件下模型高度依赖这两个具有拮抗效应的重量特征；随着λ值增大至10^{1}以上，所有特征的系数绝对值均呈现单调衰减趋势，其中壳重量的系数衰减幅度最大（从3.5降至0.2），反映其受正则化抑制最为显著。

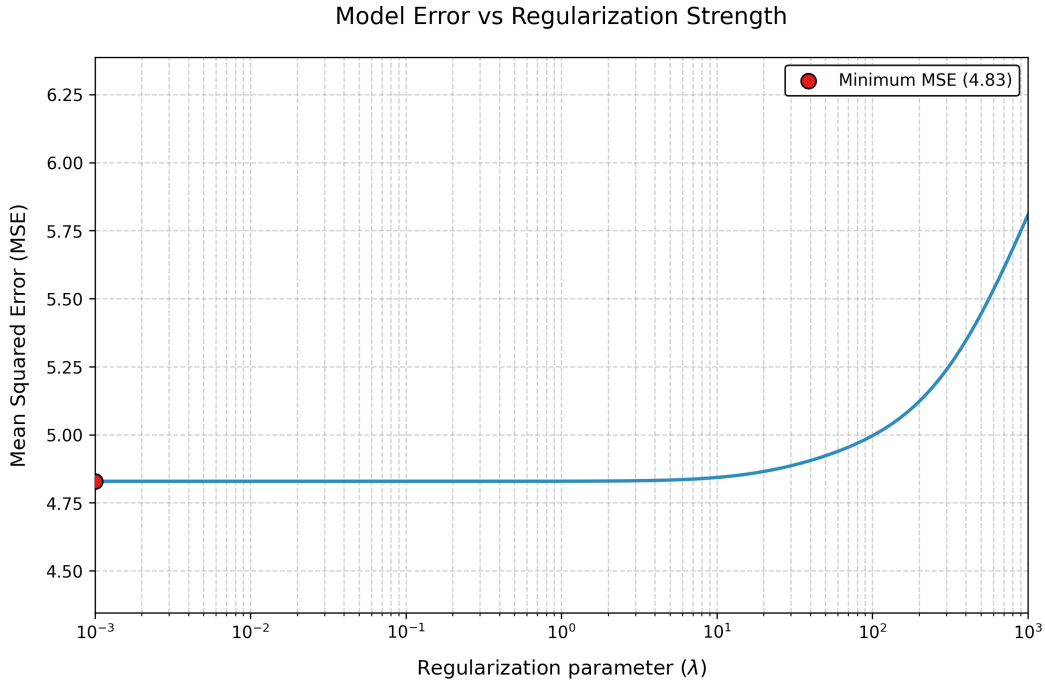

图5 相对误差随 λ 变化情况

图 5 展示了岭回归模型中正则化强度 λ 与预测误差的动态关系。

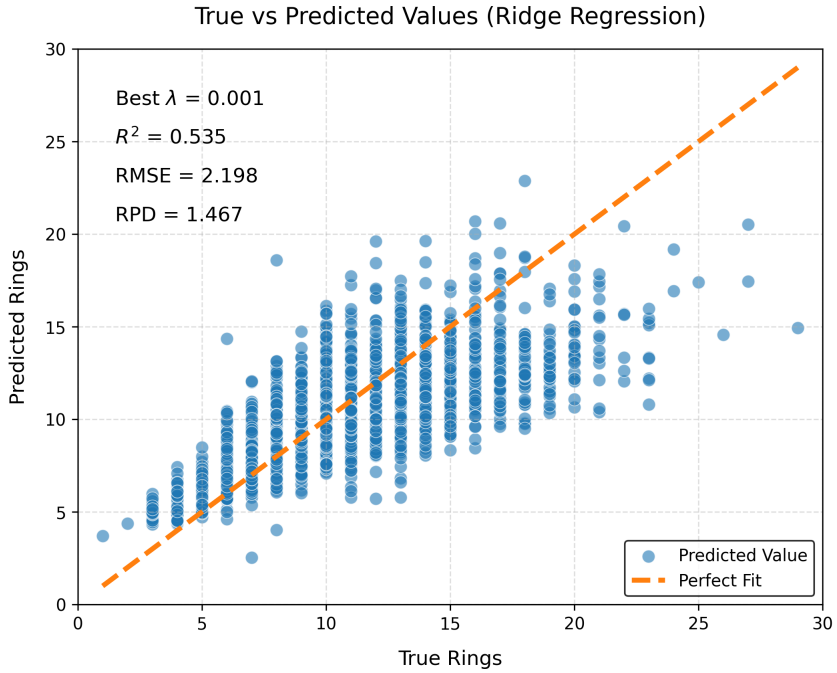

图6 预测结果散点图

图 6 为预测结果散点图，揭示了岭回归模型对鲍鱼年龄预测的系统性特征。结果表明，数据点主要集中分布在5-15环的真实值区间，在此范围内预测值呈现出以完美拟合线为中心的对称分布，但存在约±3环的离散波动；当真实值超过20环时，预测值出现明显下偏趋势（平均偏低4.2环）。

5.实验感悟

实验过程中，从数据分布分析到模型诊断的完整闭环实践，使我真正体会到"数据驱动"的含义，每一个建模决策都应建立在对数据特性的充分理解之上，而可视化正是连接数据本质与模型选择的桥梁。这种从具体问题出发，通过迭代优化寻找平衡点的思维方式，将是指导我后续研究的重要方法论。