一、数理统计初步部分

1.总体、样本、统计量的定义

2. 常见统计量：样本均值、样本方差、样本 k 阶原点矩、样本 k 阶中心矩

3. 样本均值、样本方差的均值和方差

4. 三大分布：卡方分布、t 分布，F 分布的定义，密度函数的形状，性质，分位数的概念

5. 三大分布对应的三个引理（至关重要：区间估计和假设检验要用到这三个重要的引理）

6. 参数估计的常见类型：点估计和区间估计

7. 点估计常见方法：矩估计和极大似然估计的思想、步骤和理论依据，各自的优缺点；尤其注意均匀分布的极大似然估计的特殊性

8. 区间估计的定义，单个正态分布关于均值和方差的区间估计（三种情况）包括求解步骤，每个情况对应的区间估计的公式

9. 假设检验的基本思想，可能犯的第一类错误、第二类错误，关于正态分布的均值和方差（三种情况）的假设检验的求解步骤

二、线性回归模型

1. 一元线性回归模型的定义，矩阵表示形式，基本假设，回归系数的最小二乘估计的推导过程、随机误差项方差的最小二乘估计、回归系数和随机误差项的最大似然估计推导过程、回归系数估计的区间估计、对除截距项外的回归系数的显著性检验、给定协变量值，预测因变量的值，被给出因变量的区间估计。

2. 多元线性回归模型的定义，矩阵表示形式，基本假设，回归系数的最小二乘估计的推导过程、随机误差项方差的最小二乘估计、回归系数和随机误差项的最大似然估计推导过程、回归系数估计的区间估计、对除截距项外的回归系数的显著性检验、给定协变量值，预测因变量的值，被给出因变量的区间估计。

3. 线性回归模型自由度的定义及意义。

4. 决定系数、调整的决定系数的定义，各自使用范围、区别与练习。

5. 多重共线性的判定方法有哪些？ 条件数、方差膨胀因子的定义；多重共线性的主要解决方法：岭回归，岭回归方法的主要原理，岭迹法选择调整参数。

6. 回归诊断的主要内容?

7. 普通残差、标准化残差、外学生化残差的定义，尤其注意杠杆统计量;残差图；强影响点

8. 影响分析： 异常点，杠杆统计量，DFFITS 准则，Cook 距离统计量，COVRATIO 准则，

9. 加权最小二乘方法

10. Box-Cox 变换

11.定性协变量建模

12. 计算判定系数R^2,并进行解释；线性回归模型对应的方差分析表，并用方差分析的思想进行解释（典型题目：P64-习题 3 中第 1 题）

三、重抽样方法

1. 重抽样方法的基本思想，常用重抽样方法：交叉验证法和自助法

2. 常用交叉验证方法的原理：验证集方法、留一交叉验证法、广义交叉验证法、k 拆交叉验证法

3. Bootstrap 方法的基本原理及执行步骤

四、模型选择与正则化

1. 传统子集选择方法：最优子集选择、逐步选择方法（向前逐步选择、向后逐步选择）、调整的判定系数、C_p准则、信息准则（AIC、BIC）

2. 岭回归的主要思想，目标函数及对应优化算法

3. 桥回归的主要思想，目标函数及对应优化算法

4. 惩罚变量选择方法的主要思想，好的惩罚函数的标准（无偏性、稀疏性、连续性），Lasso，SCAD，自适应的 Lasso，弹性网

五、非参数回归模型

1.非参数回归模型的定义，优缺点

2. 非参数回归模型常用估计方法：多项式回归、分段多项式拟合方法、d 阶回归样条、线性样条、三次样条、自然三次样条、样条节点个数和位置的选择（等间距方法、等间距样本分位数方法、变量选择方法）、光滑样条、局部非参数光滑方法（N-W 核光滑方法、局部多项式光滑方法）

六、Logistic 回归

1. 多元 logistic 回归模型的定义，极大似然估计法，显著性检验，预测

2. 二分类模型的评估：混淆矩阵、灵敏度、特异度、1-特异度、召回率、受试者工作特征曲线（ROC 曲线）、ROC 曲线下面积（AUC）

3. 多元 logistic 回归模型的惩罚似然变量选择方法：目标函数及优化算法

4. 非参数 logistic 回归模型的定义，估计方法

5. 多项 logistic 回归模型的定义及估计方法

《线性回归模型》重要知识点一

一、填空题（每题 2 分，共 20 分）

1. 普通最小二乘法（OLS）的目标是最小化_____。

   答案：残差平方和（SSE）

2. 在线性回归模型中，若解释变量间存在高度相关性，称为_____。

   答案：多重共线性

3. 检验回归方程整体显著性的统计量是_____。

   答案：F 统计量

4. 调整后的 R^2（Adjusted R^2）的作用是惩罚模型中_____的增加。

   答案：自变量个数

5. 异方差性会导致 OLS 估计量的_____不再有效。

   答案：标准误

6. 若残差图呈现“漏斗型”分布，表明可能存在_____。

   答案：异方差

7. 多重共线性的诊断方法包括_____和方差膨胀因子（VIF）。

   答案：相关系数矩阵

8. 回归系数的 t 检验用于检验单个自变量对_____的显著性影响。

   答案：因变量

9. 预测区间与置信区间的区别在于预测区间包含_____的不确定性。

   答案：新观测值

10. 处理异方差的常用方法是_____。

    答案：加权最小二乘法（WLS）

二、选择题（每题 2 分，共 20 分）

1. 以下哪项不是线性回归模型的经典假设？

   A. 解释变量与随机误差项不相关

   B. 随机误差项具有同方差性

   C. 随机误差项服从正态分布

   D. 解释变量之间存在多重共线性

   答案：D

2. 若回归模型的 $R^2=0.8$，则表明（ ）。

   A. 80% 的因变量变异可由自变量解释

   B. 自变量与因变量的相关系数为 0.8

   C. 回归方程的显著性水平为 0.8

   D. 残差平方和占总平方和的 80%

   答案：A

3. 多重共线性的主要影响是（ ）。

   A. 回归系数估计值不稳定

   B. 残差方差增大

   C. 模型拟合优度降低

   D. 预测精度提高

   答案：A

4. 以下哪种方法可用于检测异方差？

   A. 杜宾 - 瓦特森检验

   B. 布雷什 - 帕甘检验

   C. 方差分析

D. 卡方检验

答案：B

5. 若回归系数的 t 检验 P 值小于 0.05，则（ ）。

A. 拒绝原假设，认为该系数显著不为零

B. 接受原假设，认为该系数显著为零

C. 无法判断

D. 需进一步进行 F 检验

答案：A

6. 调整后的 R^2 与 R^2 的关系是（ ）。

A. 调整后的 $R^2 \leq R^2$

B. 调整后的 $R^2 \geq R^2$

C. 两者相等

D. 无法确定

答案：A

7. 以下哪种方法可用于处理多重共线性？

A. 增加样本量

B. 删除高度相关的自变量

C. 进行变量标准化

D. 以上都是

答案：D

8. 若残差的 Durbin-Watson 统计量为 1.2，则可能存在（ ）。

A. 正自相关

B. 负自相关

C. 异方差

D. 多重共线性

答案：A

9. 回归模型中的截距项表示（ ）。

A. 当所有自变量为零时因变量的平均值

B. 自变量的边际效应

C. 因变量的平均值

D. 模型的拟合优度

答案：A

10. 以下哪种情况会导致 OLS 估计量有偏？

A. 遗漏重要解释变量

B. 随机误差项异方差

C. 自变量间存在多重共线性

D. 随机误差项自相关

答案：A

三、计算题（每题 10 分，共 30 分）

1. **简单线性回归估计**

给定以下数据：

X	Y
1	3

X	Y
2	5
3	7
4	9
5	11

（1）计算 OLS 估计的斜率和截距。

（2）计算 R^2 值。

（3）对斜率进行 t 检验（α=0.05）。

解答：

（1）斜率 $\beta\hat{}_1=2$，截距 $\beta\hat{}_0=1$。

（2）$R^2=1$（完全拟合）。

（3）t 统计量为无穷大，拒绝原假设，斜率显著不为零。

2. **多元线性回归检验**

某回归模型结果如下：

- 总平方和（SST）= 1000
- 回归平方和（SSR）= 800
- 残差平方和（SSE）= 200
- 自变量个数 $k=3$
- 样本量 $n=50$

（1）计算 F 统计量并判断回归方程是否显著（α=0.05）。

（2）计算调整后的 R^2。

解答：

（1）$F=200/46800/3\approx61.33$，拒绝原假设，方程显著。

（2）Adjusted $R2=1-1000/49200/46\approx0.796$。

3. **置信区间计算**

某回归模型中，斜率估计值为 $\beta\hat{}_1=0.5$，标准误为 $SE(\beta\hat{}_1)=0.1$，样本量 $n=100$。

（1）计算斜率的 95% 置信区间。

（2）若要求置信区间宽度不超过 0.2，至少需要多少样本量？

解答：

（1）置信区间为 $[0.304,0.696]$。

（2）样本量需至少 $n=196$。

四、应用题（每题 15 分，共 30 分）

1. **模型诊断与改进**

某研究人员建立了一个房价预测模型，自变量包括房屋面积（X1）、房龄（X2）和周边学校数量（X3）。回归结果如下：

- 斜率估计值：$\beta\hat{}_1=200$，$\beta\hat{}_2=-50$，$\beta\hat{}_3=1000$
- 方差膨胀因子（VIF）：X1=5，X2=4，X3=10
- 残差图显示残差随 X1 增大而增大。

（1）分析模型存在的问题。

（2）提出改进建议。

解答：

（1）问题：

- X3 的 VIF=10，存在严重多重共线性。
- 残差图显示异方差性。

（2）建议：

- 删除 X3 或合并相关变量。
- 对 X1 进行变换（如对数变换）或使用加权最小二乘法处理异方差。

2. **异方差检验与处理**

某回归模型的残差平方与自变量 X 的散点图显示明显递增趋势。

（1）设计步骤检验是否存在异方差。

（2）若存在异方差，说明如何调整模型。

解答：

（1）检验步骤：

- 绘制残差平方与 X 的散点图。
- 进行 Breusch-Pagan 检验或 White 检验。

（2）处理方法：

- 对 Y 或 X 进行变换（如对数变换）。
- 使用加权最小二乘法，权重为 $1/X^2$。

《线性回归模型》重要知识点二

一、填空题（每题 2 分，共 20 分）

1. 线性回归模型的经典假设中，要求随机误差项服从_____分布。

   答案：正态

2. 若模型存在异方差，OLS 估计量仍具有_____性，但不再是有效估计量。

   答案：一致

3. 当解释变量间存在完全共线性时，回归系数的方差会趋于_____。

   答案：无穷大

4. 调整后的 R^2（Adjusted R^2）的计算公式为_____。

   答案：$1-\frac{SST/(n-1)}{SSE/(n-k-1)}$

5. 检验异方差的常用方法包括 Breusch-Pagan 检验和_____检验。

   答案：White

6. 虚拟变量陷阱是指引入虚拟变量的数量超过_____时导致的完全共线性问题。

   答案：分类数减 1

7. 回归模型中，若自变量 X 与因变量 Y 的关系随第三个变量 M 变化，则 M 称为_____变量。

   答案：调节

8. 预测区间与置信区间的区别在于预测区间包含_____的不确定性。

   答案：新观测值

9. 处理多重共线性的方法包括逐步回归、岭回归和_____。

   答案：Lasso 回归

10. 若残差的 Durbin-Watson 统计量为 0.8，则可能存在_____自相关。

答案：正

二、选择题（每题 2 分，共 20 分）

1. 以下哪项不是线性回归模型的经典假设？

   A. 解释变量与随机误差项不相关

   B. 随机误差项具有同方差性

   C. 解释变量之间存在高度相关性

   D. 随机误差项服从正态分布

   答案：C

2. 若回归模型的 $R^2=0.6$，则表明（ ）。

   A. 60% 的因变量变异可由自变量解释

   B. 自变量与因变量的相关系数为 0.6

   C. 回归方程的显著性水平为 0.6

   D. 残差平方和占总平方和的 60%

   答案：A

3. 多重共线性的主要影响是（ ）。

   A. 回归系数估计值不稳定

   B. 残差方差减小

   C. 模型拟合优度降低

   D. 预测精度提高

   答案：A

4. 以下哪种方法可用于检测异方差？

   A. 杜宾 - 瓦特森检验

   B. 布雷什 - 帕甘检验

   C. 方差分析

   D. 卡方检验

   答案：B

5. 若回归系数的 t 检验 P 值小于 0.01，则（ ）。

   A. 拒绝原假设，认为该系数显著不为零

   B. 接受原假设，认为该系数显著为零

   C. 需进一步进行 F 检验

   D. 无法判断

   答案：A

6. 虚拟变量的个数通常为分类数减 1，其目的是避免（ ）。

   A. 异方差

   B. 自相关

   C. 多重共线性

   D. 模型设定偏误

   答案：C

7. 以下哪种方法可用于处理异方差？

   A. 加权最小二乘法

   B. 逐步回归法

   C. 主成分分析

D. 方差分析

答案：A

8. 若残差图呈现“漏斗型”分布，表明可能存在（ ）。

A. 异方差

B. 自相关

C. 多重共线性

D. 模型设定错误

答案：A

9. 回归模型中的交互项表示（ ）。

A. 自变量对因变量的独立影响

B. 自变量之间的相互影响

C. 因变量对自变量的影响

D. 模型的非线性关系

答案：B

10. 以下哪种情况会导致 OLS 估计量有偏？

A. 遗漏重要解释变量

B. 随机误差项异方差

C. 自变量间存在多重共线性

D. 随机误差项自相关

答案：A

三、计算题（每题 10 分，共 30 分）

1. **简单线性回归估计与检验**

   给定以下数据：

X	Y
1	3
2	5
3	7
4	9
5	11

2.

3. （1）计算 OLS 估计的斜率和截距。

   （2）计算 R^2 值并解释其含义。

   （3）对斜率进行 t 检验（$\alpha=0.05$）。

4. **解答：**

   （1）斜率 $\beta\hat{}_1=2$，截距 $\beta\hat{}_0=1$。

   （2）$R^2=1$，表明因变量的变异完全由自变量解释。

   （3）t 统计量为无穷大，拒绝原假设，斜率显著不为零。

5. **多元线性回归模型诊断**

   某回归模型结果如下：

   - 总平方和（SST）= 2000
   - 回归平方和（SSR）= 1500
   - 残差平方和（SSE）= 500
   - 自变量个数 $k=4$

- 样本量 $n=100$

（1）计算 F 统计量并判断回归方程是否显著（α=0.05）。

（2）计算调整后的 R^2。

（3）若某自变量的方差膨胀因子（VIF）= 8，分析模型存在的问题。

解答：

（1）$F=500/951500/4\approx71.25$，拒绝原假设，方程显著。

（2）Adjusted $R2=1-2000/99500/95\approx0.753$。

（3）VIF=8 表明存在严重多重共线性，需删除或合并相关变量。

6. **非线性回归模型转换**

某研究发现因变量 Y 与自变量 X 存在非线性关系，数据如下：

X	Y
1	2
2	5
3	10
4	17
5	26

（1）建议对 Y 或 X 进行何种转换以线性化模型。

（2）转换后计算 OLS 估计的斜率和截距。

（3）验证转换后的模型是否满足线性假设。

解答：

（1）对 Y 取对数或对 X 进行平方转换。

（2）若转换为 $\ln(Y)$，斜率 $\beta\hat{}_1 \approx 0.693$，截距 $\beta\hat{}_0 \approx 0$。

（3）绘制残差图，若残差随机分布则满足线性假设。

四、应用题（每题 15 分，共 30 分）

1. **模型诊断与改进**

某研究人员建立了一个员工薪资预测模型，自变量包括工作年限（X1）、学历（X2，虚拟变量：本科 = 0，硕士 = 1）和绩效评分（X3）。回归结果如下：

- 斜率估计值：$\beta\hat{}_1 = 500$，$\beta\hat{}_2 = 2000$，$\beta\hat{}_3 = 300$
- 方差膨胀因子（VIF）：X1=2，X2=1.5，X3=10
- 残差图显示残差随 X3 增大而增大。

（1）分析模型存在的问题。

（2）提出改进建议。

解答：

（1）问题：

- X3 的 VIF=10，存在严重多重共线性。
- 残差图显示异方差性。

（2）建议：

- 删除 X3 或使用岭回归处理多重共线性。
- 对 X3 进行对数变换或使用加权最小二乘法处理异方差。

2. **虚拟变量与交互项应用**

某电商平台分析广告投入（X1）和促销活动（X2，虚拟变量：无促销 = 0，有促销 = 1）对销售额（Y）的影响。回归模型为：$Y=\beta_0+\beta_1X_1+\beta_2X_2+\beta_3X_1X_2+\epsilon$

回归结果如下：

- β^0=1000，β^1=20，β^2=500，β^3=10
- 促销活动组的 R^2=0.85，非促销组的 R^2=0.65

（1）解释各回归系数的经济意义。

（2）比较促销与非促销时广告投入对销售额的边际效应。

（3）分析模型的拟合效果差异。

解答：

（1）β_1=20：非促销时，广告投入每增加 1 元，销售额增加 20 元；β_2=500：促销时，基础销售额增加 500 元；β_3=10：促销时，广告投入的边际效应额外增加 10 元。

（2）促销时边际效应为 20+10=30 元，非促销时为 20 元。

（3）促销组 R^2 更高，说明促销活动显著提升模型拟合效果。

《模型评价》重要知识点

一、填空题（每题 2 分，共 20 分）

1. 回归模型中，衡量预测值与真实值绝对差异的指标是_____。

答案：平均绝对误差（MAE）

2. 分类模型中，准确率的计算公式为_____。

   答案：(TP+TN)/(TP+TN+FP+FN)

3. 交叉验证中，将数据集划分为 K 个互不重叠子集的方法称为

   _____。

   答案：K 折交叉验证

4. 信息准则 AIC 的计算公式为_____。

   答案：-2ln (L) + 2k（L 为似然函数，k 为参数数量）

5. 当模型在训练集上表现良好但在测试集上表现差时，称为_____。

   答案：过拟合

6. 多分类问题中，将所有类别预测结果汇总计算整体 F1 值的方法称为

   _____。

   答案：微平均 F1（Micro-F1）

7. 模型解释性方法中，基于博弈论的加性解释框架是_____。

   答案：SHAP（Shapley Additive exPlanations）

8. 偏差 - 方差分解中，方差反映模型对_____的敏感程度。

   答案：训练数据变化

9. 检验两个分类变量关联性的统计方法是_____。

   答案：卡方检验

10. 处理数据不平衡问题时，常用的评估指标是_____。

    答案：AUC-ROC 或 F1 值

二、选择题（每题 2 分，共 20 分）

1. 以下哪项不是回归模型的评估指标？

   A. 均方根误差（RMSE）

B. 决定系数（R^2）

C. 准确率（Accuracy）

D. 平均绝对百分比误差（MAPE）

答案：C

2. 当数据存在严重类别不平衡时，以下哪种指标更可靠？

A. 准确率

B. 召回率

C. F1 值

D. 以上都是

答案：C

3. 交叉验证的主要作用是（ ）。

A. 减少过拟合

B. 提高模型复杂度

C. 增加训练数据量

D. 以上都不是

答案：A

4. AIC 与 BIC 的主要区别在于（ ）。

A. BIC 对模型复杂度的惩罚更严厉

B. AIC 适用于大样本数据

C. BIC 更关注模型拟合度

D. AIC 包含样本量信息

答案：A

5. 以下哪种方法可用于检测模型的过拟合？

   A. 绘制学习曲线

   B. 计算方差膨胀因子（VIF）

   C. 布雷什 - 帕甘检验

   D. 杜宾 - 瓦特森检验

   答案：A

6. 多分类问题中，宏平均 F1（Macro-F1）的计算方式是（ ）。

   A. 对每个类别计算 F1 值后取平均

   B. 对所有类别计算全局 F1 值

   C. 对每个样本计算 F1 值后取平均

   D. 以上都不是

   答案：A

7. 以下哪种方法可用于模型解释性分析？

   A. 特征重要性分析

   B. 混淆矩阵

   C. 交叉验证

   D. 方差分析

   答案：A

8. 偏差 - 方差分解中，高偏差通常对应（ ）。

   A. 欠拟合

   B. 过拟合

   C. 模型复杂度高

D. 数据量不足

答案：A

9. 检验两个独立模型性能差异的统计方法是（ ）。

A. 卡方检验

B. 双样本 t 检验

C. 方差分析

D. 费舍尔精确检验

答案：B

10. 以下哪种方法可用于处理异方差问题？

A. 加权最小二乘法

B. 岭回归

C. Lasso 回归

D. 逐步回归

答案：A

三、计算题（每题 10 分，共 30 分）

1. **回归模型评估指标计算**

给定以下数据：

真实值 y_i	**预测值** $\hat{y}_i$
5	4.8
3	3.2

真实值 y_i	预测值 $\hat{y}_i$
7	6.5
6	6.1
2	2.3

（1）计算 MAE、RMSE 和 R^2。

（2）若真实值的均值为 4.6，分析模型的拟合效果。

解答：

（1）MAE = 0.32，RMSE ≈ 0.36，R^2 ≈ 0.985。

（2）R^2 接近 1，说明模型拟合效果极佳。

2. **分类模型评估指标计算**

某二分类模型的混淆矩阵如下：

	预测正例	预测反例
真实正例	80	20
真实反例	15	85

（1）计算准确率、精确率、召回率和 F1 值。

（2）若正样本占比 20%，分析模型的性能。

解答：

（1）准确率 = 0.825，精确率≈0.842，召回率 = 0.8，F1≈0.821。

（2）准确率较高但正样本召回率偏低，需关注正样本识别能力。

3. **多分类模型评估指标计算**

某三分类模型的混淆矩阵如下：

	预测 A	预测 B	预测 C
真实 A	90	5	5
真实 B	2	85	13
真实 C	1	10	89

（1）计算宏平均 F1 和微平均 F1。

（2）若类别 A 样本量远大于 B 和 C，分析哪种指标更合理。

解答：

（1）宏平均 F1≈0.88，微平均 F1≈0.88。

（2）微平均 F1 更合理，因其受大类样本主导。

四、应用题（每题 15 分，共 30 分）

1. **模型诊断与改进**

某研究人员建立了一个房价预测模型，回归结果如下：

- R^2=0.95，调整后 R^2=0.94
- MAE=5000 元，RMSE=7000 元
- 学习曲线显示训练误差和验证误差均随数据量增加而下降，但仍存在差距。

（1）分析模型存在的问题。

（2）提出改进建议。

解答：

（1）问题：

- 高 R^2 但 MAE 和 RMSE 较大，可能存在异常值影响。
- 学习曲线显示模型仍有欠拟合空间。

（2）建议：

- 检查数据中的异常值并处理。
- 增加模型复杂度（如使用多项式回归）或调整超参数。

2. **模型解释性与偏差 - 方差分析**

某分类模型在训练集上准确率为 95%，测试集上准确率为 70%，且特征重要性显示模型过度依赖某一特征。

（1）分析模型的偏差 - 方差情况。

（2）使用 SHAP 值解释模型决策过程。

（3）提出改进建议。

解答：

（1）高方差导致过拟合，模型对训练数据敏感。

（2）SHAP 值可显示每个特征对预测结果的贡献，识别过度依赖的特征。

（3）建议：

- 增加正则化（如 L2 正则）。
- 进行特征选择或降维。
- 采用集成学习（如随机森林）降低方差。

《重抽样方法》重要知识点

一、填空题

1. **Bootstrap** 的核心思想是通过_____从原始数据中生成多个自助样本。

   答案：有放回抽样

2. **Jackknife** 方法通过每次删除_____个观测值来估计统计量的偏差和方差。

   答案：一

3. 交叉验证中，将数据集划分为训练集和测试集的重复次数称为_____。

   答案：折数（或 K）

4. **Bootstrap 置信区间**的计算方法包括正态近似法、枢轴量法和_____。

   答案：分位数法

5. 处理数据不平衡问题时，可在 Bootstrap 中对少数类进行_____以提高模型敏感性。

   答案：过采样

6. 当原始数据存在异常值时，_____方法对统计量的稳定性评估更可靠。

   答案：Jackknife

7. **留一交叉验证**（LOOCV）的样本量为 n 时，需进行_____次模型训练。

   答案：n

8. **双重 Bootstrap** 可用于估计统计量的_____，即 Bootstrap 估计本身的不确定性。

   答案：标准误差

9. 自助聚集（Bagging） 通过并行训练多个模型来降低_____，提升预测稳定性。

   答案：方差

10. 若 Bootstrap 样本中某观测值未被选中的概率为_____，则其被称为“袋外数据”。

    答案：$(1-1/n)_n \approx 1/e$

二、选择题

1. 以下哪项不是 Bootstrap 的特点？

   A. 有放回抽样

   B. 适用于小样本

   C. 依赖正态分布假设

   D. 可估计置信区间

   答案：C

2. Jackknife 方法主要用于（ ）。

   A. 估计统计量的偏差

   B. 处理异方差

   C. 特征选择

   D. 模型预测

   答案：A

3. 以下哪种方法可用于检测异常值对统计量的影响？

   A. 自助法（Bootstrap）

   B. 刀切法（Jackknife）

   C. 交叉验证

   D. 方差分析

   答案：B

4. 当数据存在严重类别不平衡时，以下哪种重抽样方法更有效？

   A. 简单随机抽样

   B. 分层 Bootstrap

   C. 留一交叉验证

   D. 系统抽样

   答案：B

5. 以下哪种方法可用于模型性能的无偏估计？

   A. 自助法（Bootstrap）

   B. 刀切法（Jackknife）

   C. 交叉验证

   D. 以上都是

   答案：D

6. Bootstrap 置信区间的分位数法直接基于（ ）。

   A. 自助统计量的分布

   B. 正态分布假设

   C. 卡方分布

D. 学生 t 分布

答案：A

7. 以下哪种方法可用于处理多重共线性？

A. 岭回归 Bootstrap

B. 随机森林

C. 刀切法

D. 以上都是

答案：D

8. 当原始数据量 $n=10$ 时，以下哪种方法计算量最大？

A. 5 折交叉验证

B. 留一交叉验证

C. 自助法（1000 次迭代）

D. 刀切法

答案：B

9. 以下哪种方法可用于估计模型参数的标准差？

A. Bootstrap

B. Jackknife

C. 交叉验证

D. 以上都是

答案：D

10. 若 Bootstrap 置信区间包含原假设值，则（ ）。

A. 拒绝原假设

B. 接受原假设

C. 无法判断

D. 需进一步检验

答案：C

三、计算题

1. **Bootstrap 均值估计与置信区间计算**

给定以下样本数据：$X=\{2,4,6,8,10\}$

（1）计算样本均值 $\bar{X}$。

（2）通过自助法（B=5 次迭代）生成 5 个自助样本，并计算对应的均值。

（3）使用分位数法计算 95% 置信区间。

解答：

（1）$\bar{X}=6$

（2）自助样本示例：

- 样本 1：[2,4,6,8,10] → 均值 = 6
- 样本 2：[4,4,8,8,10] → 均值 = 6.8
- 样本 3：[2,6,6,10,10] → 均值 = 6.8
- 样本 4：[4,6,6,8,10] → 均值 = 6.8
- 样本 5：[2,2,4,8,10] → 均值 = 5.6

（3）排序后的自助均值：5.6, 6, 6.8, 6.8, 6.8

95% 置信区间为$[5.6,6.8]$。

2. **Jackknife 方差估计**

某研究团队使用 Jackknife 方法估计回归系数的方差。原始样本量 $n=5$，每次删除一个观测值后得到的回归系数估计值为：$\beta\hat{}_{-1}=2.1$, $\beta\hat{}_{-2}=1.9$, $\beta\hat{}_{-3}=2.0$, $\beta\hat{}_{-4}=2.2$, $\beta\hat{}_{-5}=1.8$。

（1）计算 Jackknife 估计值 $\beta\hat{}_{\text{jack}}$。

（2）计算方差估计 $\text{Var}(\beta\hat{}_{\text{jack}})$。

解答：

（1）$\beta\hat{}_{\text{jack}}=5\times\bar{\beta}-4\times\bar{\beta}_{-i}$

原始估计值 $\bar{\beta}=(2.1+1.9+2.0+2.2+1.8)/5=2.0\beta\hat{}_{\text{jack}}$

$=5\times2.0-4\times2.0=2.0$

（2）方差估计：$\text{Var}=54\times\sum i=15(\beta\hat{}-i-\beta\bar{}-i)2\beta\bar{}-i=2.0\sum(\beta\hat{}-i$

$-2.0)2=(0.1)2+(-0.1)2+02+0.22+(-0.2)2=0.1\text{Var}=54\times0.1=0.08$

3. **交叉验证模型评估**

某分类模型在 10 折交叉验证中的准确率如下：

$$[0.82,0.85,0.81,0.83,0.84,0.86,0.80,0.87,0.83,0.84]$$

（1）计算平均准确率和标准差。

（2）若原假设为平均准确率 $\mu=0.8$，使用 t 检验判断是否显著高于原假设（α=0.05）。

解答：

（1）平均准确率 = 0.835，标准差≈0.024

（2）t 统计量 = $0.024/100.835-0.8\approx4.64$

自由度 = 9，临界值≈1.833，拒绝原假设，显著高于 0.8。

四、应用题

1. **模型稳定性评估与改进**

某研究人员建立了一个房价预测模型，使用 Bootstrap 方法评估模型参数的稳定性。结果显示：

- 回归系数 $\beta\hat{}1$ 的 Bootstrap 均值为 1200，标准差为 150
- 回归系数 $\beta\hat{}2$ 的 Bootstrap 均值为 800，标准差为 300
- 模型 R^2 的 Bootstrap 均值为 0.75，标准差为 0.05

（1）分析模型存在的问题。

（2）提出改进建议。

解答：

（1）问题：

- β^2 的标准差较大，说明参数估计不稳定。
- R^2 的标准差较小，表明模型整体拟合效果稳定，但个别参数波动大。

（2）建议：

- 检查自变量 X_2 是否存在异常值或多重共线性。
- 增加正则化（如岭回归）以降低参数方差。
- 扩大样本量或使用更稳健的估计方法（如 M 估计）。

2. **数据不平衡与重抽样策略**

某医疗数据集包含 90% 的健康样本和 10% 的患病样本，使用 Logistic 回归进行分类。

（1）分析直接使用原始数据训练模型的潜在问题。

（2）设计三种重抽样策略并比较其优缺点。

（3）若要求模型对患病样本的召回率≥0.9，应选择哪种策略？

解答：

（1）问题：模型易偏向多数类，患病样本识别率低。

（2）策略：

- **过采样少数类**：增加患病样本数量，但可能导致过拟合。
- **欠采样多数类**：减少健康样本数量，可能损失信息。
- **分层 Bootstrap**：在每次抽样中保持类比例，平衡训练集。

（3）选择**过采样少数类**或**分层 Bootstrap**，以确保患病样本被充分学习。

《模型选择与正则化》重要知识点

一、填空题

1. AIC（赤池信息准则）的计算公式为_____。

答案：$\mathrm{AIC}=2k-2\ln(L)$（k 为参数数量，L 为似然函数）

2. **Lasso 回归**的正则化项为_____。

   答案：$\lambda \sum_{j=1}^{p} | \beta_j |$

3. 当模型在训练集和测试集上均表现不佳时，称为_____。

   答案：欠拟合

4. **交叉验证**中，将数据集划分为训练集和验证集的重复次数称为_____。

   答案：折数（或 K）

5. ** 弹性网（Elastic Net）** 的正则化项结合了_____和_____。

   答案：L1 正则化、L2 正则化

6. **BIC（贝叶斯信息准则）** 对模型复杂度的惩罚比 AIC 更_____。

   答案：严厉

7. 岭回归通过引入_____来防止过拟合。

   答案：L2 正则化项

8. 模型选择中，**自由度**指模型中_____的数量。

   答案：可自由调整的参数

9. **偏差 - 方差权衡**中，高方差通常对应_____。

   答案：过拟合

10. 处理高维数据时，**Lasso 回归**的优势是_____。

    答案：自动进行特征选择

二、选择题

1. 以下哪项不是模型选择的准则？

   A. AIC

   B. R^2

   C. BIC

   D. 交叉验证

   答案：B

2. 当样本量 n 较大时，BIC 更倾向于选择（ ）。

   A. 复杂模型

B. 简单模型

C. 中等复杂度模型

D. 以上都不是

答案：B

3. 以下哪种方法可用于特征选择？

A. 岭回归

B. Lasso 回归

C. 弹性网

D. B 和 C

答案：D

4. 正则化参数 λ 越大，模型的（ ）。

A. 方差越小，偏差越大

B. 方差越大，偏差越小

C. 方差和偏差均减小

D. 方差和偏差均增大

答案：A

5. 交叉验证的主要作用是（ ）。

A. 减少过拟合

B. 提高模型复杂度

C. 估计模型泛化误差

D. 增加训练数据量

答案：C

6. 以下哪种方法可用于处理多重共线性？

A. 岭回归

B. Lasso 回归

C. 逐步回归

D. 以上都是

答案：A

7. 当模型存在高方差时，应（ ）。

   A. 增加正则化强度

   B. 减少正则化强度

   C. 增加模型复杂度

   D. 减少样本量

   答案：A

8. 弹性网的正则化项为（ ）。

   A. $\lambda\sum|\beta_j|$

   B. $\lambda\sum\beta_{j2}$

   C. $\lambda_1\sum|\beta_j|+\lambda_2\sum\beta_{j2}$

   D. $\lambda(\sum|\beta_j|+\sum\beta_{j2})$

   答案：C

9. 以下哪种方法可用于模型解释性分析?

   A. 特征重要性分析

   B. 混淆矩阵

   C. 交叉验证

   D. 方差分析

   答案：A

10. 若两个模型的 AIC 值相同，应选择（ ）。

    A. 参数数量较多的模型

    B. 参数数量较少的模型

    C. 无法判断

    D. 需进一步比较 BIC

    答案：D

三、计算题

1. **AIC 与 BIC 计算**

   某线性回归模型包含 3 个参数，残差平方和（RSS）为 100，样本量 $n=50$。

（1）计算 AIC 和 BIC。

（2）若另一模型的 AIC=120，BIC=130，哪个模型更优？

解答：

（1）$AIC=2\times3+50\ln(\frac{50}{100})\approx6+50\times0.693\approx40.65BIC=3\ln(50)+50\ln(\frac{50}{100})\approx3\times3.912+50\times0.693\approx11.736+34.65\approx46.386$

（2）原模型 AIC 和 BIC 均更小，更优。

2. **岭回归参数估计**

给定数据集 $X=[1324]$，$y=[57]$，正则化参数 $\lambda=1$。

（1）计算岭回归的参数估计 $\beta^{\wedge}$。

（2）比较岭回归与普通最小二乘法（OLS）的参数估计差异。

解答：

（1）$\beta^{\wedge}=(X^{T}X+\lambda I)^{-1}X^{T}yX^{T}X=[10141420],X^{T}y=[2638]X^{T}X+\lambda I=[11141421]$

解得 $\beta^{\wedge}=[1.02.0]$。

（2）OLS 估计为 $[-1.03.0]$，岭回归通过正则化缩小了参数值。

3. **交叉验证模型选择**

某分类模型在 5 折交叉验证中的准确率为：$[0.78,0.82,0.79,0.81,0.80]$。

（1）计算平均准确率和标准差。

（2）若原假设为平均准确率 $\mu=0.8$，使用 t 检验判断是否显著高于原假设（$\alpha=0.05$）。

解答：

（1）平均准确率 = 0.80，标准差≈0.0158。

（2）t 统计量 $= 0.0158/50.80-0.8\approx 0$，不拒绝原假设。

四、应用题

1. **模型选择与正则化应用**

   某研究团队建立了一个房价预测模型，使用 Lasso 回归进行特征选择。结果显示：

   - 模型在训练集上的 $R^2=0.95$，测试集上的 $R^2=0.75$
   - 部分特征的系数被压缩为零
   - 学习曲线显示训练误差和验证误差均随数据量增加而下降，但仍存在差距

   （1）分析模型存在的问题。

   （2）提出改进建议。

   解答：

   （1）问题：

   - 测试集 R^2 较低，存在过拟合。
   - 学习曲线显示模型仍有欠拟合空间，可能正则化强度不足。

   （2）建议：

   - 调整 Lasso 的正则化参数 λ，增加惩罚力度。
   - 检查数据中的异常值并处理。
   - 尝试弹性网结合 L1 和 L2 正则化。

2. **高维数据与正则化策略**

   某基因表达数据集包含 10000 个特征和 500 个样本，使用 Logistic 回归进行疾病分类。

   （1）分析直接使用普通 Logistic 回归的潜在问题。

   （2）设计三种正则化策略并比较其优缺点。

   （3）若要求模型同时实现特征选择和参数收缩，应选择哪种策略？

   解答：

（1）问题：

- 高维数据易导致过拟合。
- 普通 Logistic 回归参数估计不稳定。

（2）策略：

- **Lasso 回归**：通过 L1 正则化实现特征选择，但可能忽略相关特征。
- **岭回归**：通过 L2 正则化收缩参数，但不进行特征选择。
- **弹性网**：结合 L1 和 L2，平衡特征选择和参数收缩。

（3）选择**弹性网**，因其能同时处理特征选择和参数收缩。

《非参数回归模型》重要知识点

一、填空题

1. **核回归**的核心思想是通过**局部加权平均**对每个点进行拟合，权重由**核函数**和**带宽**决定。

   答案：局部加权平均、核函数、带宽
2. **局部多项式回归**通过在每个点附近拟合**低阶多项式**来捕捉非线性关系，其权重计算基于数据点的**距离**。

   答案：低阶多项式、距离
3. **样条回归**通过分段多项式在**节点**处保持连续性和光滑性，其中**自然样条**在边界处强制为**线性**以减少边界效应。

   答案：节点、自然样条、线性
4. **广义加性模型（GAM）的结构是将响应变量建模为多个平滑函数**的和，通常选择**样条函数**作为平滑基函数。

   答案：平滑函数、样条函数
5. **交叉验证**常用于选择非参数回归中的**带宽**或**节点数量**，以平衡模型的偏差和方差。

   答案：带宽、节点数量
6. **核函数**的选择影响局部加权回归的平滑程度，常见类型包括**高斯核**和**Epanechnikov 核**。

   答案：高斯核、Epanechnikov 核

7. **非参数回归**的评估指标包括**均方误差（MSE）**、**决定系数（R^2）和交叉验证误差**。

   答案：均方误差（MSE）、决定系数（R^2）

8. **样条回归**的自由度由**节点数量**和**多项式次数**共同决定，例如三次样条的自由度为**K+4**（K 为节点数）。

   答案：节点数量、多项式次数、K+4

9. **局部加权散点平滑（LOWESS）是一种改进的局部多项式回归方法，通过迭代加权**提高对异常值的鲁棒性。

   答案：迭代加权

10. **非参数回归**的主要缺点包括**不能外推**、**小样本效果差**和**高维诅咒**。

    答案：不能外推、小样本效果差

二、选择题

1. 以下哪项不是核回归的特点？

   A. 局部加权平均

   B. 依赖数据驱动

   C. 需要假设模型形式

   D. 适用于非线性关系

   答案：C

2. 局部多项式回归与核回归的主要区别在于（ ）。

   A. 是否使用核函数

   B. 是否拟合多项式

   C. 是否处理高维数据

   D. 是否需要带宽

   答案：B

3. 样条回归的自由度由（ ）决定。

   A. 节点数量

   B. 多项式次数

   C. 数据量

D. A 和 B

答案：D

4. GAM 的平滑函数通常选择（ ）。

   A. 样条函数

   B. 核函数

   C. 线性函数

   D. 以上都是

   答案：A

5. 以下哪种方法可用于处理非参数回归中的边界效应？

   A. 镜像反射法

   B. 交叉验证

   C. 正则化

   D. 分层抽样

   答案：A

6. 当数据存在明显非线性趋势时，以下哪种方法更有效？

   A. 核回归

   B. 线性回归

   C. 岭回归

   D. Lasso 回归

   答案：A

7. 非参数回归的缺点包括（ ）。

   A. 不能外推

   B. 小样本效果差

   C. 高维诅咒

   D. 以上都是

   答案：D

8. 以下哪种方法可用于估计非参数回归模型的标准差？

   A. Bootstrap

   B. Jackknife

C. 交叉验证

D. 以上都是

答案：D

9. 局部多项式回归的优点包括（ ）。

A. 减少边界效应

B. 自适应拟合

C. 计算效率高

D. A 和 B

答案：D

10. 若数据存在严重噪声，以下哪种方法更鲁棒？

A. 核回归

B. 局部多项式回归

C. 样条回归

D. 线性回归

答案：B

三、计算题

1. **核回归预测值计算**

给定数据集 $X=\{1,3,5,7,9\}$，$Y=\{2,4,6,8,10\}$，使用高斯核函数 $K(x)=h2\pi1 e-2h2(x-xi)2$，带宽 $h=2$，计算 $x=6$ 处的预测值。

解答：

（1）计算每个数据点的核权重：$K(6,1)=22\pi1e-2\times22(6-1)2\approx0.003K(6,3)=22\pi1 e-2\times22(6-3)2\approx0.054K(6,5)=22\pi1e-2\times22(6-5)2\approx0.199K(6,7)=22\pi1e-2\times22(6-7)2 \approx0.199K(6,9)=22\pi1e-2\times22(6-9)2\approx0.054$

（2）归一化权重：$\sum K=0.003+0.054+0.199+0.199+0.054=0.509wi =0.509Ki$

（3）预测值：$y^{\wedge}=\sum w_i Y_i=0.5090.003\times2+0.054\times4+0.199\times6+0.199\times8+0.054\times10\approx7.0$

2. **样条回归自由度计算**

某自然三次样条模型包含 5 个节点，计算其自由度。

解答：

自然三次样条的自由度为节点数 + 多项式次数 - 边界约束数。

多项式次数为 3，边界约束数为 2（两端线性），因此：$自由度=5+3-2=6$

3. **交叉验证选择带宽**

某核回归模型在 5 折交叉验证中的均方误差（MSE）如下：$h=1$: 2.1$h=2$: 1.8$h=3$: 2.3$h=4$: 2.5$h=5$: 2.7

（1）选择最优带宽。

（2）若原假设为最优带宽 $h=2$，使用 t 检验判断是否显著优于 $h=1$（α=0.05）。

解答：

（1）最优带宽为 $h=2$（MSE 最小）。

（2）t 统计量 $= \frac{5-1(2.1-1.8)2+(1.8-1.8)2+(2.3-1.8)2+(2.5-1.8)2+(2.7-1.8)2}{1.8-2.1}\approx-1.23$

自由度 = 4，临界值≈2.776，不拒绝原假设，无显著差异。

四、应用题

1. **非参数回归模型选择与比较**

某研究团队分析房价数据，比较核回归、局部多项式回归和样条回归的拟合效果：

- 核回归在带宽 $h=3$ 时，训练集 MSE=1.2，测试集 MSE=1.5
- 局部多项式回归在多项式阶数 = 2 时，训练集 MSE=1.0，测试集 MSE=1.8
- 样条回归在节点数 = 5 时，训练集 MSE=0.8，测试集 MSE=2.0

（1）分析各模型的表现。

（2）提出改进建议。

解答：

（1）表现分析：

- 核回归：测试集 MSE 较低，模型复杂度适中。
- 局部多项式回归：训练集过拟合，测试集 MSE 较高。
- 样条回归：训练集过拟合，测试集 MSE 最高。

（2）建议：

- 局部多项式回归：降低多项式阶数或增加正则化。
- 样条回归：减少节点数或使用自然样条。
- 核回归：尝试不同带宽或结合交叉验证优化。

2. **GAM 在医学数据中的应用**

某医疗研究使用 GAM 分析患者年龄与血压的关系，模型结构为：$血压=\beta_0+f_1(年龄)+f_2(性别)+\epsilon$

其中 f_1 为自然样条函数，f_2 为指示函数。

（1）解释模型的意义。

（2）若年龄的平滑函数显示非线性趋势，如何调整模型？

（3）若性别系数不显著，应如何处理？

解答：

（1）模型意义：血压由年龄的非线性函数和性别差异共同解释。

（2）调整方法：增加年龄的节点数或使用更高阶的样条函数。

（3）处理方法：删除性别变量或使用更简单的参数模型。

《Logistic 回归》重要知识点

一、填空题

1. **Logistic 回归**的核心思想是通过 **logit 变换**将线性回归模型的输出映射到（0,1）区间，用于预测**二分类**事件的概率。

   答案：logit 变换、二分类

2. **极大似然估计（MLE）是 Logistic 回归参数估计的常用方法，其目标是最大化对数似然函数**的值。

   答案：对数似然函数

3. **优势比（OR）表示自变量每增加一个单位时，事件发生的优势变化倍数，其计算公式为 exp(β)**。

   答案：exp(β)

4. **Hosmer-Lemeshow 检验**用于评估 Logistic 回归模型的**拟合优度**，原假设为**模型预测值与观测值一致**。

   答案：拟合优度、模型预测值与观测值一致

5. 哑变量（Dummy Variable）**用于处理分类自变量，若某分类变量有 k 个类别，则需创建 $k-1$** 个哑变量。

   答案：$k-1$

6. **多重共线性**会导致参数估计不稳定，常用 方差膨胀因子（VIF）作为诊断指标，当 VIF>**10** 时提示存在严重共线性。

   答案：方差膨胀因子（VIF）、10

7. **正则化**可用于处理过拟合问题，Logistic 回归中常用的正则化方法包括 **L1** 和 **L2** 正则化。

   答案：L1、L2

8. **ROC 曲线**的横坐标为**假阳性率（FPR）**，纵坐标为**真阳性率（TPR）**，曲线下面积（AUC）越接近 1 表示模型性能越好。

   答案：假阳性率（FPR）、真阳性率（TPR）

9. **有序 Logistic 回归**适用于因变量为**有序分类变量**的场景，需满足**比例优势假设**。

   答案：有序分类变量、比例优势假设

10. **逐步回归**是变量筛选的常用方法，包括**向前法**、**向后法**和**逐步法**。

    答案：向前法、向后法

二、选择题

1. 以下哪项不是 Logistic 回归的应用场景？

   A. 疾病风险预测

   B. 客户流失分析

   C. 房价趋势预测

   D. 文本情感分类

   答案：C

2. Logistic 回归的因变量必须是（ ）。

   A. 连续变量

   B. 二分类变量

   C. 多分类变量

   D. 有序分类变量

   答案：B

3. 以下哪种方法可用于处理分类自变量？

   A. 标准化

   B. 归一化

   C. 哑变量编码

   D. 主成分分析

   答案：C

4. 若模型存在严重多重共线性，应采取的措施是（ ）。

   A. 增加样本量

   B. 删除相关变量

   C. 提高正则化参数

   D. 以上都是

   答案：D

5. 以下哪个指标用于评估模型的预测能力？

   A. 对数似然值

   B. AIC

   C. AUC-ROC

D. 偏差

答案：C

6. 有序 Logistic 回归需满足的关键假设是（ ）。

A. 正态性

B. 独立性

C. 比例优势

D. 方差齐性

答案：C

7. 正则化参数 λ 越大，模型的（ ）。

A. 方差越小，偏差越大

B. 方差越大，偏差越小

C. 方差和偏差均减小

D. 方差和偏差均增大

答案：A

8. 以下哪种方法可用于模型变量筛选？

A. 逐步回归

B. 岭回归

C. Lasso 回归

D. 以上都是

答案：D

9. 若某分类变量有 4 个类别，需创建（ ）个哑变量。

A. 1

B. 2

C. 3

D. 4

答案：C

10. Hosmer-Lemeshow 检验的原假设是（ ）。

A. 模型拟合不佳

B. 模型预测值与观测值一致

C. 自变量与因变量无关

D. 存在多重共线性

答案：B

三、计算题

1. **Logistic 回归参数估计与概率计算**

   给定数据集 $X=135246$，$Y=011$，使用极大似然估计求解 Logistic 回归模型参数 β_0,β_1,β_2，并计算 $x=(2,3)$时的预测概率。

   解答：

   （1）构建似然函数：L(\beta) = \prod_{i=1}^n \frac{1}{1+e^{-(\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2})}}}

   （2）取对数并求导，解得参数估计：$\beta_0\approx-2.5,\beta_1\approx1.0,\beta_2\approx0.5$

   （3）预测概率：$p=1+e^{-(-2.5+1.0\times2+0.5\times3)}1\approx0.62$

2. **优势比计算与解释**

   某 Logistic 回归模型中，自变量“吸烟”的回归系数 $\beta=0.8$，计算其优势比（OR）并解释含义。

   解答：$\mathrm{OR}=e^{0.8}\approx2.225$

   解释：吸烟者患病的优势是非吸烟者的 2.225 倍。

3. **Hosmer-Lemeshow 检验**

   某模型的 Hosmer-Lemeshow 检验统计量为 $\chi^2=8.5$，自由度为 8，判断模型拟合优度（α=0.05）。

   解答：

   查表得临界值 $\chi^2_{0.05,8}=15.51$，因 8.5 < 15.51，**不拒绝原假设**，模型拟合良好。

四、应用题

1. **Logistic 回归模型构建与诊断**

某研究分析糖尿病患病风险，自变量包括年龄、BMI、家族史（0 = 无，1 = 有），因变量为是否患病（0 = 否，1 = 是）。模型结果如下：

- 年龄系数 $\beta=0.05$（p<0.01）
- BMI 系数 $\beta=0.12$（p=0.03）
- 家族史系数 $\beta=1.2$（p<0.01）
- 模型 AUC=0.85，Hosmer-Lemeshow 检验 p=0.15

（1）解释各变量的影响。

（2）评估模型性能。

（3）若存在多重共线性，如何处理？

解答：

（1）影响分析：

- 年龄每增加 1 岁，患病优势增加 $e^{0.05}\approx1.051$ 倍。
- BMI 每增加 1 单位，患病优势增加 $e^{0.12}\approx1.127$ 倍。
- 有家族史者患病优势是无家族史者的 $e^{1.2}\approx3.32$ 倍。

（2）性能评估：

- AUC=0.85，模型预测能力较好。
- Hosmer-Lemeshow p=0.15>0.05，模型拟合良好。

（3）处理共线性：

- 计算 VIF 值，若 VIF>10，剔除相关变量或使用正则化。

2. **多分类 Logistic 回归应用**

某电商平台分析用户购买偏好，因变量为手机品牌（苹果、华为、小米），自变量包括价格、屏幕尺寸、电池续航。

（1）选择合适的 Logistic 回归类型。

（2）解释模型参数的意义。

（3）若价格系数为 - 0.02（p<0.01），如何解读？

解答：

（1）类型：**多分类 Logistic 回归**（因变量为无序多分类）。

（2）参数意义：

- 以苹果为参照类，价格每增加 1 元，选择华为的优势变化为 $e^{\beta_{\text{华为-价格}}}$。
- 同理可解释其他变量对小米的影响。

（3）解读：

- 价格每增加 1 元，用户选择华为的优势降低 $e^{-0.02} \approx 0.98$ 倍，即价格越高，用户越倾向于选择其他品牌。