绿色表示第一个选项的条件，黄色表示第二个选项的条件。对于 CARS 数据集，还输入了一个额外的二进制变量 "昂贵（便宜）"，如果汽车价格高于 35000（低于 20000），该变量的值为真。

1) 对主要预测因子和分层预测因子的组合提出自己的变式，以满足以下条件：在卡方检验中，响应对主要预测因子的依赖具有统计学意义（显著性水平 0.05），而在有第二个预测因子参 与的分层 CMH 检验中，对原始预测因子的依赖不再具有显著性。连续预测因子的离散化版本可以作为预测因子（建议不要做太多的抽样间隔，在大多数情况下，对连续预测因子进行二进制分割就足够了）。
2) 使用阶跃函数建立廉价（昂贵）逻辑回归，使用正向（反向）方法选择变量，并列举模型的所有 "复杂性"，从一个变量到所有变量（反之亦然）。对于搜索得到的每个模型，使用 ROC 曲线下面积（又称一致性统计量）进行交叉验证（5 个区块），评估其质量，绘制质量得分（CV ROC AUC）与模型复杂度（预测因子数量）的关系图，并选择最佳模型（基本上确定最佳模型的变量列表）。
3) 在整个样本和超采样样本（比例为 1:1）上训练最佳模型，在整个样本上为这两个模型建立引导 ROC 曲线（您可以自己编写代码或使用 boot.roc 函数）。超采样模型的质量是否有明显变化?