题目一

某家保险公司希望将车辆的安全性与其他几个变量联系起来。他们使用保险索赔的频率作为基础，给每个车型一个分数。 数据保存在Safety数据集中。

Safety数据集中包含如下变量：

Unsafe 安全分数(低于平均水平为1，与平均水平持平或高于平均水平为0)

Type 车型(Large, Medium, Small, Sport/Utility, Sports)

Region 产地(Asia, N America)

Weight 车重(千磅)

Size 对Type进行三分，Small/Sports为1，Medium为2，Large为3

使用Unsafe作为响应变量，Weight作预测变量，对低于平均安全分数(Unsafe=1)的概率建模，拟合Logistic回归模型，使用剖面似然(Profile likelihood)方法计算优比(Odds Ratio)的置信区间，绘制优比图(Odds ratio plots)和效应图(Effect plots)，并回答以下问题：

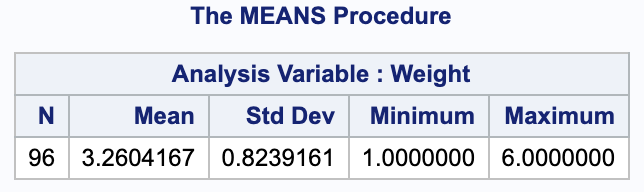
1. 是否拒绝回归系数为0的原假设？
2. 写出Logistic回归方程。
3. 解释Weight优比的意义。

【答】

一、建模准备

在正式开始建模前，首先对给出的数据进行描述性分析。针对此题，对Weight变量做描述性分析，得到样本容量、均值、标准差、最大值和最小值等描述量。

具体结果如下：



经分析，对于本例，可以使用Unsafe作为响应变量，Weight作预测变量，对低于平均安全分数(Unsafe=1)的概率建模，并拟合Logistic回归模型。

二、建立模型

记低于平均安全分数的概率为，对应地，与平均水平持平或高于平均水平的概率为，有。

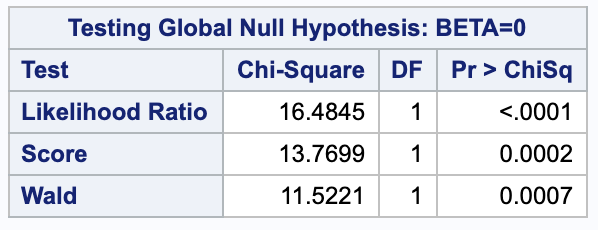
建立logit模型如下

需要编写SAS程序估计上式中的和两个值。

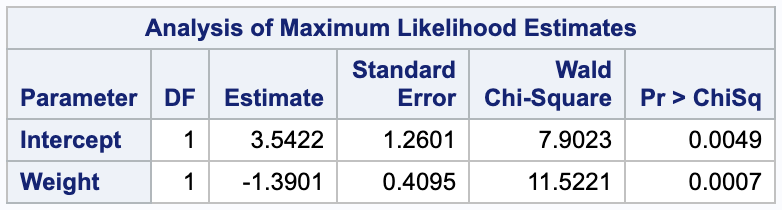
SAS程序源代码见附件中./solution.sas

三、问题解答

1. 观察回归系数为0的假设，ChiSq观测值均小于0.05，说明在95%的置信水平上，Weight自变量具有显著性，因此拟合出来的模型显著优于仅含有常数项的模型。即，拒绝回归系数为0的原假设。

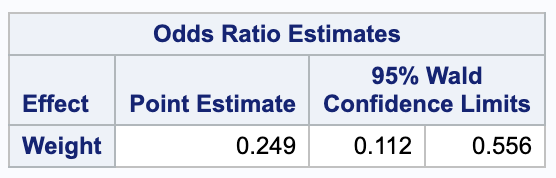


2. 观察下表进行读数，得出回归方程为：

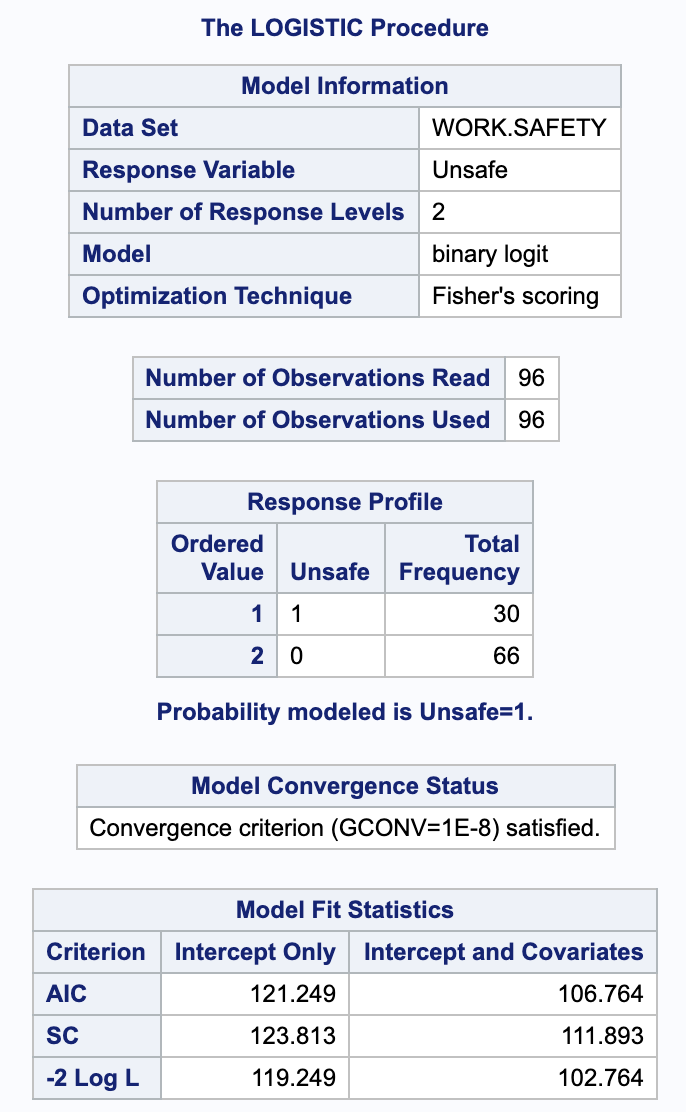


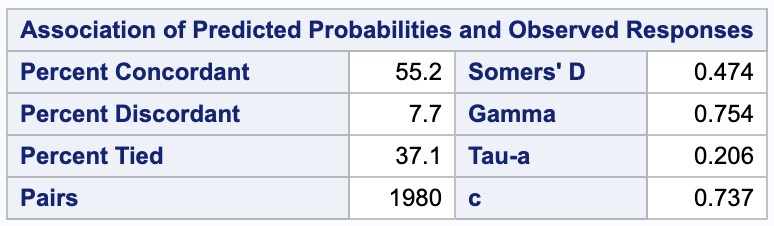
3. 观察优比表格，读数得出以下结果：

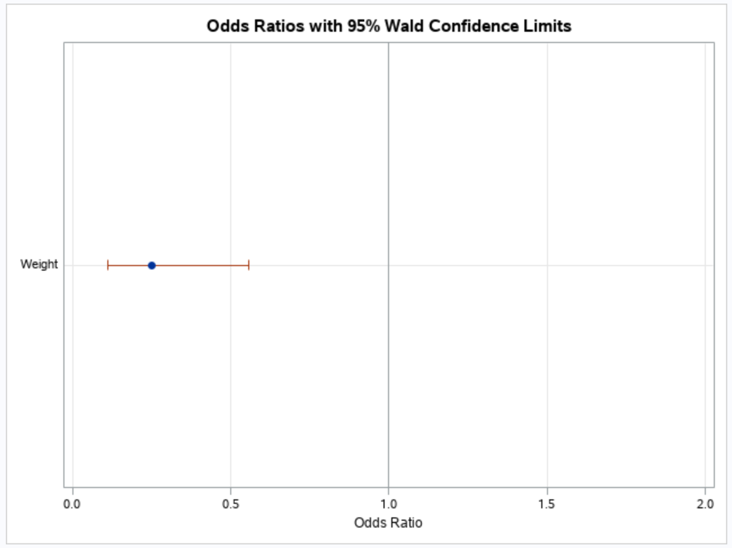
* Weight优比的意义为，每当自变量Weight变化一个单位，胜率（本题中为Unsafe为1的概率与为0的概率之比）的变化值
* 具体为，当Weight每增加1，回导致下降1.3901，取对数后，胜率将变成原来的倍
* 即，随着Weight的增加，车辆的安全性系数就越高，出现低于平均安全分数的概率就越低

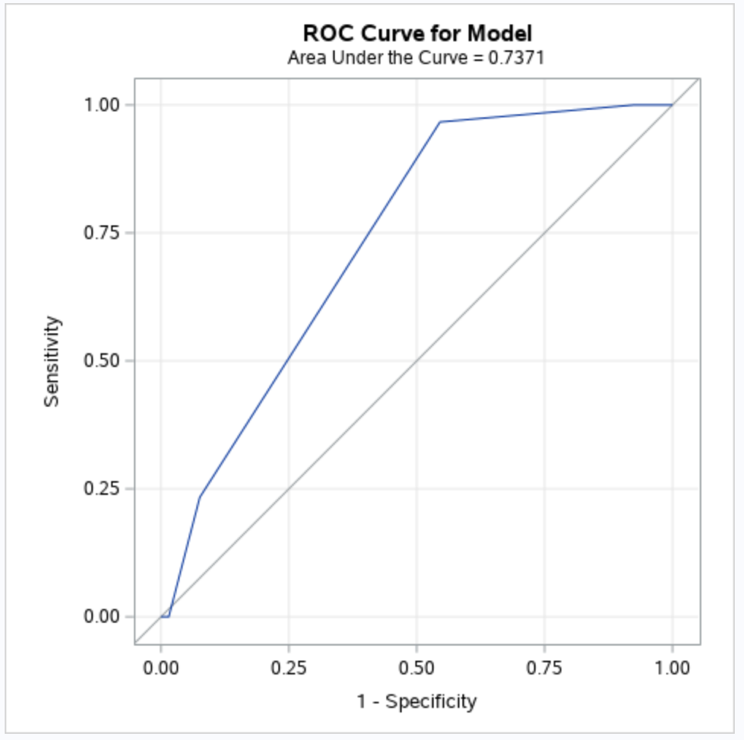


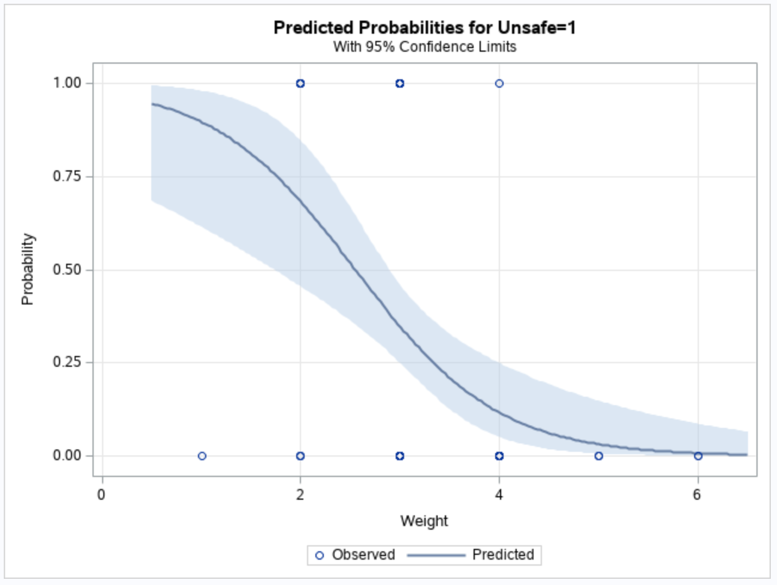
四、其他运行结果展示











题目二

【答】选择 C。

在线性回归中，是用直线去拟合数据，实现最小二乘意义下的最小预测误差。在逻辑回归中：，可以看作是用直线去拟合Logit函数，通过极大似然估计出参数，使得在该参数下，能以最大概率生成当前的样本。这里要说明的是，线性回归解决的是回归问题，而逻辑回归是分类问题，但两者的形式非常的相似，上面两式的右边也是一致的。

Logistic回归通过对数据分类边界的拟合来实现分类。而这条数据分类边界即为直线，这也是它为什么可以被看作是一个广义线性模型的原因。在线性回归中，回归的因变量y是连续的，没有明确的上下限，因此可以用线性模型来拟合。而逻辑回归应用于分类时，因变量，也就是类别y只有0和1，满足二项分布，这是连续的线性模型无法拟合的。因此，需要选择最佳的连接函数，它就是Logit函数。Logit函数能把自变量从(0,1)连续单调地映射到正负无穷，这里类别y的0和1值分别对应(0,0.5)和(0.5,1)的概率值p。另外，把看作一个整体，反解出p，就会看到我们熟悉的sigmoid函数。

郑重声明

**本作业由作者独立完成。抄袭行为在任何情况下都是不能容忍的(COPY is strictly prohibited under any circumstances)！由抄袭所产生的一切后果由抄袭者承担，勿谓言之不预也。**

**陈麒先**