北京大学
PEKING UNIVERSITY

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：

The MEANS Procedure

Analysis Variable : Weight				
N	Mean	Std Dev	Minimum	Maximum
96	3.2604167	0.8239161	1.0000000	6.0000000

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$p1 = P(unsafe = 1)$，对应地，与平均水平持平或高于平均水平的概率为$p0 = P(unsafe = 0)$，有$p0 + p1 = 1$。

建立 logit 模型如下

$$logit = \ln\left(\frac{p1}{p0}\right) = \beta_0 + \beta_1 \cdot Weight$$

需要编写 SAS 程序估计上式中的β_0和β_1两个值。

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。

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	16.4845	1	<.0001
Score	13.7699	1	0.0002
Wald	11.5221	1	0.0007

2. 观察下表进行读数，得出回归方程为：

$$logit = \ln\left(\frac{p_1}{p_0}\right) = 3.5422 - 1.3901 \cdot Weight$$

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	3.5422	1.2601	7.9023	0.0049
Weight	1	-1.3901	0.4095	11.5221	0.0007

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$\ln\left(\frac{p_1}{p_0}\right)$下降 1.3901，取对数后，胜率$\left(\frac{p_1}{p_0}\right)$将变成原来的$e^{-1.3901} = 0.249$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

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
Weight	0.249	0.112	0.556

四、其他运行结果展示

The LOGISTIC Procedure

Model Information	
Data Set	WORK.SAFETY
Response Variable	Unsafe
Number of Response Levels	2
Model	binary logit
Optimization Technique	Fisher's scoring

Number of Observations Read	96
Number of Observations Used	96

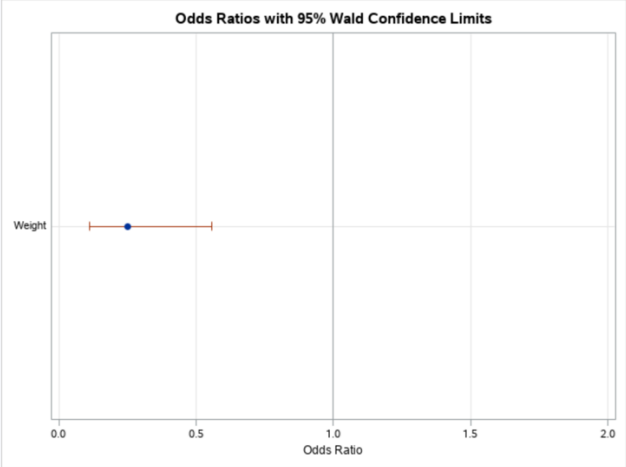
Response Profile		
Ordered Value	Unsafe	Total Frequency
1	1	30
2	0	66

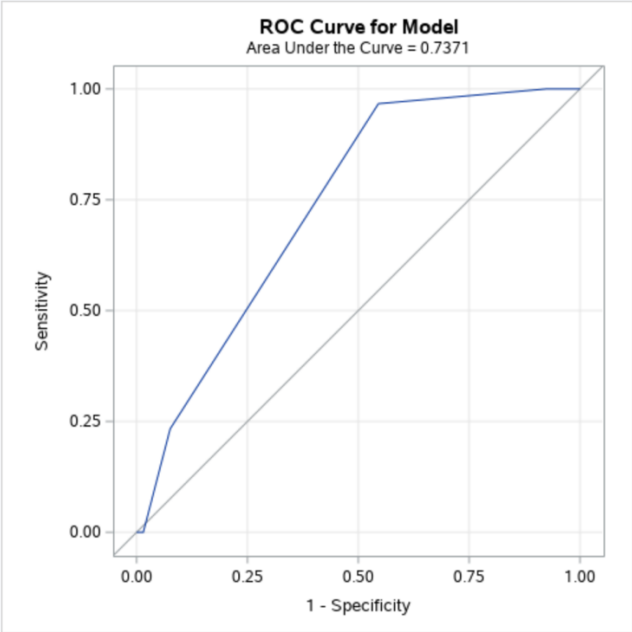
Probability modeled is Unsafe=1.

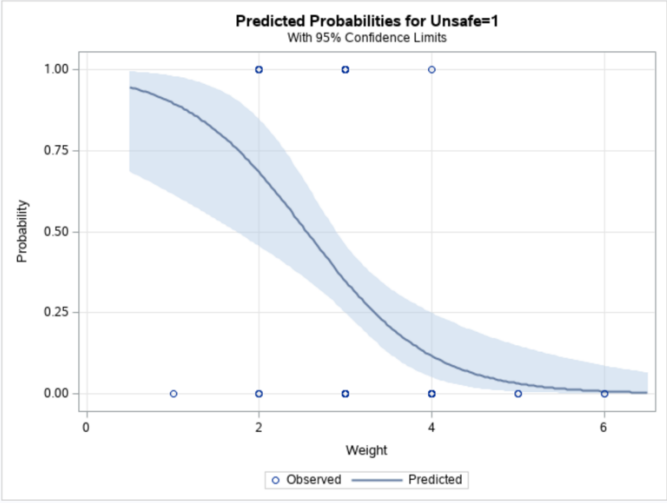
Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	121.249	106.764
SC	123.813	111.893
-2 Log L	119.249	102.764

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	55.2	Somers' D	0.474
Percent Discordant	7.7	Gamma	0.754
Percent Tied	37.1	Tau-a	0.206
Pairs	1980	c	0.737







题目二

【答】选择 C。

在线性回归中，$y = w^T x$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：$logit(p) = logit\left(\frac{p}{1-p}\right) = w^T x$，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$w^T x$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