肝癌手术治疗效果评价

摘　要： 本文主要研究了某医院 10 年来肝癌病例中的 20 个有代表性的样本，通过建立逻辑回归的数据挖掘模型对预后有影响的 10 个指标进行分析，实现对预后效果的预测以及寻找各个变量之间的重要程度，从而为病人规划最佳的手术治疗方案。在建模中，我们首先通过对样本数据进行基本的描述性统计分析，初步观测每个指标对预后影响是否显著。然后再利用逻辑回归模型对预后影响进行预测。进一步的，为了消除指标中的共线性性以及寻找对预后有显著影响的指标，我们在原方法的基础上采用了逐步回归的思想优化原有的逻辑回归模型，通过设置显著性水平的阈值来发现对预后有重要影响的指标，并且达到消除相关性的目的。得出了是否有食道静脉曲张、HbsAg 和 Anti-HCV 的阴阳性、肿瘤大小、肿瘤的包膜否完整以及肿瘤旁的微小子灶这 5 个指标都对预后具有很大影响的结论。最后，我们利用舍一法以及 ROC 曲线来比较是否有优化的两个模型预测效果，得到 ROC 曲线线下两个模型面积分别为 0.78020，和 0.6538，表明分类预测效果良好。.

关键词：数据挖掘 逻辑回归 逐步回归 ROC 曲线

Liver cancer Surgical treatment evaluation

Abstract: This paper mainly studied 20 representative sample cases of liver cancer in a hospital in the past 10 years ,in the paper, we establishment the logistic regression model to analyze the prognostic impact of 10 indicators, in order to predict the prognosis and find the important ones among them, so that we can plan the best surgical treatment options for patients.In the model,we firstly use the basic descriptive statistical analysis and preliminary observations for each indicator whether it has a significant influence on the prognosis.Then we use the logistic regression model to predict the impact on prognosis.Futher,in order to eliminate the collinearity of the indexs as well as looking for the indexs which have a significant impact on prognosis.We use the stepwise regression method to optimize the existing model,by setting the significant level of threshold ,we eliminate the collinearity of the indexs and conclude that “whether esophageal varices”, “the negative or positive of HBsAg and anti-HCV”, “tumor size”, “the tiny sub stove next to the tumor” and “whether the tumor capsule is complete” these five indexs have a high impact on the prognosis.Finally we use the “Give up one Method” and the ROC curve to compare the effect of the two models which has optimized or not and The area under the ROC curve line of them is 0.78020 and 0.6538,respectively,indicating that the classification and prediction effect is good.

Key words: Data mining; logistic regression; stepwise regression; ROC curve

目　录

1. 挖掘目标

本次建模主要针对某医院 10 年来肝癌病例中的 20 个有代表性的样本，选取对预后有影响的 10 个指标进行统计分析；以预后影响作为评价标准，建立数据挖掘模型，实现对手术的治疗效果的自动分类和方案的优劣进行预测，从而为病人规划最佳的手术和治疗方案。

2. 分析方法与过程

2.1. 总体流程

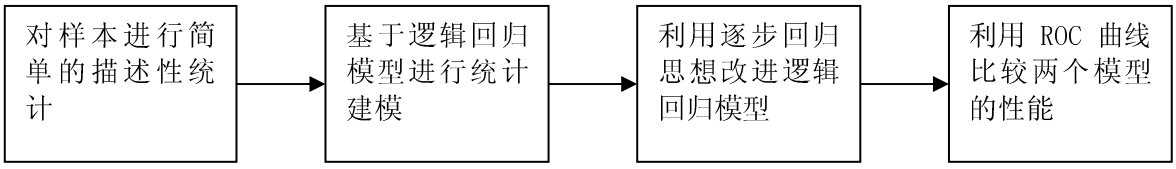
为了让建模更为清晰，结合该 20 个样本的特点，我们建模的主要步骤如下：

一、 针对本数据集的特点，对该样本进行简单的描述性统计，并设计出指标变量；

二、 基于逻辑回归模型的统计建模，实现对手术的治疗效果的自动分类和方案的优劣进行预测，并对模型结果给出合理的解释；

三、 利用逐步回归思想改进逻辑回归模型，并进行两个模型进行比较模型优良。

四、 基于 ROC 曲线比较以上两种分类器的性能，给出最优模型。



图一 建模主要步骤流程图

2.2. 具体步骤

2.2.1 数据介绍

在详细介绍建模之前，我们给出数据集如下

表 1 预处理后样本数据

X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	DECISION
mid	branch	negative	negative	rightliver	middle	dilation	part	no	less	Y
mid	trunk	positive	positive	rightliver	middle	infiltration	no	have	much	N
serious	no	negative	positive	leftliver	big	dilation	no	no	much	Y
no	no	negative	negative	allliver	verybig	dilation	integrate	no	much	Y
light	branch	positive	positive	rightliver	small	infiltration	integrate	have	no	N
mid	trunk	positive	negative	rightliver	middle	infiltration	part	no	no	Y
light	branch	positive	negative	rightliver	small	infiltration	no	have	much	Y
no	trunk	negative	positive	allliver	big	dilation	part	no	less	N
mid	branch	positive	negative	rightliver	middle	dilation	integrate	have	less	N

续上表

X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	DECISION
no	no	negative	positive	rightliver	verybig	dilation	part	no	no	N
serious	trunk	negative	negative	rightliver	big	infiltration	integrate	have	less	N
light	trunk	positive	negative	allliver	small	dilation	no	no	less	Y
no	no	negative	positive	rightliver	verybig	infiltration	part	no	much	N
no	branch	negative	negative	allliver	verybig	infiltration	integrate	have	no	N
serious	branch	positive	positive	rightliver	big	infiltration	part	have	less	Y
mid	no	negative	positive	rightliver	middle	dilation	integrate	have	much	N
mid	branch	negative	negative	rightliver	middle	dilation	integrate	no	less	N
light	trunk	negative	positive	leftliver	small	infiltration	no	no	no	Y
no	trunk	negative	positive	rightliver	verybig	infiltration	no	no	no	N
no	no	positive	positive	rightliver	verybig	dilation	part	no	less	Y

表 2 指标变量说明

变量名	变量标示	变量说明
X1	食道静脉曲张	无(no)、轻(light)、中(mid)、重(serious)
X2	门脉癌栓	无(no)、分支(branch)、主干(trunk)
X3	HbsAg	阴性(negative)、阳性(positive)
X4	Anti-HCV	阴性(negative)、阳性(positive)
X5	肿瘤部位	左肝(leftliver)、右肝(rightliver)、全肝(allliver)
X6	肿瘤大小	直径<3cm(small)、3~5cm(middle)、5~10cm(big)、>10cm(verybig)
X7	肿瘤生长方式	膨胀(dilation)、浸润(infiltration)
X8	肿瘤包膜	完整(integrate)、子灶突破包膜(part)、无(no)
X9	肿瘤旁的微小子灶	无(no)、有(have)
X10	术后腹水	无(no)、少(less)、多(much)
DECISION	预后影响	有(Y)、无(N)

这里一共有 10 个变量指标，其中 X1 到 X10 为解释变量，DECISION 为被解释变量。在这里除了 X6 可以转化为数值型变量外，其余都是定性变量指标，我们不能使用简单的回归进行建模，必须考虑定性变量的性质。从以上表中可以看到，这 20 个样本中 9 个预后有影响、11 个预后没有影响的样本，为了更清晰明了地了解预后影响和其它变量的关系，我们需要进行初步的描述性统计分析。

2.2.2 描述性统计

本建模应该考虑哪些指标变量呢？换句话说，哪些解释变量会对被解释变量 DECISION 有较大影响呢？如何对 X1 到 X10 这些变量进行预处理，以转化为可分析的指标变量？以下我们以预后影响 DECISION 为 Y 时设计为 1，为 N 时设计为 0 来表识是否有预后影响。并以此为因变量对各解释变量进行描述性统计，以便找出各解释变量的进一步转化。

1.食道静脉曲张（X1）

表 3 食道静脉曲张的描述性统计

食道静脉曲张	no	light	mid	serious
总数	7	4	6	3
预后有影响	29%	75%	33%	67%
预后没有影响	71%	25%	67%	33%

那么该如何解读解读这个表格呢？就拿第二列来说吧，7 表示 20 个样本中有 7 个样本 X1 变量的值是 no，其中 7 个 X1 变量的值为 no 的样本中有约 29%是预后有影响的，其余的可以类似的来解释。但是我们发现有轻微的食道静脉曲张的样本中，预后有影响的占较大比例，这也超过中度以及严重程度时的比例，这也许是样本量太少造成的误差，为此我们可以考虑对 X1 有无食道静脉曲张来进行分类，重新统计，我们可得有食道静脉曲张的样本量为 13 个，其中有 7 个对预后有影响，占比为 54%，这远大于没有食道静脉曲张的 29%。从数据出发，我们可得，没有食道静脉曲张的患者具有更好的预后效果。

2. 门脉癌栓(X2)

表 4 门脉癌栓的描述性统计

门脉癌栓(X2)	no	branch	trunk
总数	7	7	6
预后有影响比例	43%	43%	50%
预后没有影响比例	57%	57%	50%

从上表可以看出，门脉癌栓在三个不同属性下对预后影响的比例并没有显著性差异，但由于在临床实践中发现，肝癌门静脉栓的形成是影响肝癌预后的重要因素，临床发生率高达 60%-90%，可惜的是迄今为止肝癌门静脉栓形成的原因尚不明确。

3. HbsAg(X3)与 Anti-HCV (X4)

表 5 HbsAg 与 Anti-HCV 的描述性统计

	Negative (X3)	Positive (X3)	Negative (X4)	Positive (X4)
总数	12	8	9	11
预后有影响	33%	62.5%	56%	36%
预后没有影响	67%	37.5%	44%	64%

阳性 HbsAg 相对于阴性 HbsAg 对预后影响具有显著差异性，而且从表中可以看出，相比于阴性 HbsAg，阳性 HbsAg 且预后有影响占有更大的比例，这说明 HbsAg 为阴性的肝癌患者具有更好的预后效果。同样 Anti-HCV 的阳性和阴性也对预后的影响有很大不同，这个差异也是相对明显的，可以看出这个变量很大可能对预后具有较大影响，同时可见 Anti-HCV 阳性患者的预后效果相比于阴性患者的效果更好。可惜的是，Anti-HCV 的阳性，即丙性肝炎病毒抗体阳性说明患者曾经感染或者正在感染丙型肝炎，这对预后会有不良影响，这也许是数据量太少，造成这种统计上的偏差。在考虑建模时需要特别注意该变量。

4.肿瘤部位（X5）

表 6 肿瘤部位的描述性统计

肿瘤部位	rightliver	leftliver	allliver
总数	14	2	4
预后有影响	36%	100%	50%
预后没有影响	64%	0	50%

我们直观的感觉是，左右肝都有肿瘤的话预后影响的概率也会大点，而只有左肝或右肝有肿瘤预后有影响应该会更小，经过再次统计，我们也发现发现左右肝都有肿瘤的对预后有影响（50%）比只有左肝或右肝有影响（0.44）稍大。

5. 肿瘤大小（X6）

表 7 肿瘤大小的描述性统计

肿瘤大小	small	middle	big	verybig
总数	4	6	4	6
预后有影响	75%	33%	50%	33%
预后没有影响	25%	67%	50%	67%

从初步数据看来，肿瘤大小对预后影响并没有很明显的结论，有可能这是一个并不是很重要的指标，其影响相对较小。由于这个变量是具有数值上意义的，我们可以用它们的中位数或者平均值代替其各水平的值，直径<3cm(small)、3~5cm(middle)、5~10cm(big)、>10cm(verybig)分别用 x6 等于 1.5、4、7.5 以及 10 来数值化该变量。

6. 肿瘤生长方式（X7）与肿瘤的包膜（X8）

表 8 肿瘤生长方式与肿瘤的包膜的描述性统计

	Infiltration（X7）	Dilation（X7）	Integrate（X8）	part（X8）	no（X8）
总数	10	10	7	7	6
预后有影响	40%	50%	14%	57%	67%
预后没有影响	60%	50%	86%	43%	33%

从肿瘤生长方式可以看出，浸润和膨胀两者的总数相同，而却两者中预后有影响的比例相差不大。膨胀性生长，肿瘤向周围扩散，挤压周围组织或邻近器官。周围可形成纤维性包膜。浸润性生长，瘤细胞沿组织间隙或毛细淋巴管扩展。一般而言，浸润式生长的肿瘤会更恶性。但对于肿瘤的包膜而言，肿瘤的包膜是完整的样本中，预后有影响所占的比例（14%）远小于其他两种情况。

7. 肿瘤旁的微小子灶（X9）与术后腹水（X10）

表 9 肿瘤旁的微小子灶与术后腹水的描述性统计

	have (X9)	no (X9)	no (X10)	less (X10)	much (X10)
总数	8	12	6	8	6
预后有影响	25%	58%	33%	50%	50%
预后没有影响	75%	42%	67%	50%	50%

从肿瘤旁的微小子灶上看，有微小子灶的患者明显比有微小子灶的患者预后好，这与我们的经验有冲突，作为预测的话，我们需要特别注意这个变量。而术后是否有腹水方面来看，没有腹水的患者更倾向于具有预后影响。

2.2.3 指标设计

在描述分析的基础上，我们对模型中需要用的的指标重新设计，具体如下表：

表 10 重新设计的指标变量说明

变量名	变量解释
x11	没有食道静脉曲张取值为 1，否则为 0
x12	轻度食道曲张取值为 1，否则为
X13	中度食道曲张取值为 1，否则为 0
x21	没有门脉癌栓取值为 1，否则为 0
x22	branch 门脉癌栓为 1，否则为 0
x3	如果 HbsAg 为阳性则记 x3 为 1，阴性记为 0
x4	若果 Anti-HCV 为阳性为 1，否则为 0
x51	肿瘤位置在左肝脏为 1，否则为 0
x52	肿瘤位置在右肝脏为 1，否则为 0
x6	肿瘤大小，数值型变量
x7	肿瘤生长方式，如果整个肝脏设为 1，否则为 0
x81	肿瘤的包膜完整为 1，否则为 0
x82	肿瘤的包膜部分为 1，否则为 0
x9	肿瘤旁的微小子灶，没有微小子灶设为 0，有则为 1
x101	术后腹水，没有腹水为 1，否则为 0
x102	术后腹水，少量腹水为 1，否则为 0

2.2.4 统计模型

虽然描述性统计能在一定程度上给我们一些信息，但是由于我们考虑的时候都是单独考虑的，并没有从整体出发，忽略了各变量之间的相关关系，这难免会造成不准确，所以我们仍然需要进行系统的统计建模，把所有的变量放在一起考虑，以降低分析的失误。

在指标设计以及描述统计的基础上，讨论如何建立回归模型。我们关心的问题是，哪些指标可能影响预后影响。由于我们希望根据各解释变量的情况预测出最后的预后是否有影响，这是一个很经典的分类问题。对于这一分类问题的建模，我们可以采用贝叶斯分类、决策树分类、随机森林法、支持向量机以及逻辑回归，它们各有各的优缺点，在这里我们主要给出逻辑回归模型，并基于这一模型给出相应的结论。

逻辑回归分析是用来处理分类问题的一种统计建模方法，我们可以建立如下的逻辑回归模型:

$$p(X'\beta)=\frac{e^{X'\beta}}{1+e^{X'\beta}}$$

或者等价地有

$$logit\{p(X'\beta)\}=log\{\frac{p(X'\beta)}{1-p(X'\beta)}\}=X'\beta$$

这里

$$X'\beta=\beta_0+\beta_1X_1+\beta_{21}X_{21}+\beta_{22}X_{22}+\beta_3X_3+\cdots++\beta_{10}X_{10}$$

这就是我们需要建立的逻辑回归模型。

同普通线性回归模型相似，对于逻辑回归而言，人们关心回归系数 β 。对于一个给定的变量 X_j ，$\beta_j=0$ 意味着在给定其他解释变量不变的前提下，该指标对于解释条件概率 $p(X'\beta)$ 没有任何帮助。因此对于解释因变量 Y 的随机行为也没有任何帮助。但是，如果 $\beta_j>0$ ，那么在给定其他解释变量不变的前提下，指标 X_j 的上升会带来条件概率 $p(X'\beta)$ 的上升，也就是说，因变量取值为 1 的可能性会变大。从某种角度看来，这似乎是一种“正相关”。 如果 $\beta_j<0$ ，那么在给定其他解释变量不变的前提下，指标 X_j 的上升会带来条件概率 $p(X'\beta)$ 的下降，也就是说，因变量取值为 0 的可能性会变大。从某种角度看来，这似乎是一种“负相关”。

考虑到变量间的共线性影响，而且并不一定所有的变量对因变量都有很好的解释作用，为了进一步优化模型，我们考虑选择逐步回归的方法进行变量选择，再次进行逻辑回归得到我们的结果。为此如下给出全模型以及优化模型的结果，并比较其作为分类器的优劣。

2.3. 结果分析

2.3.1 模型结果

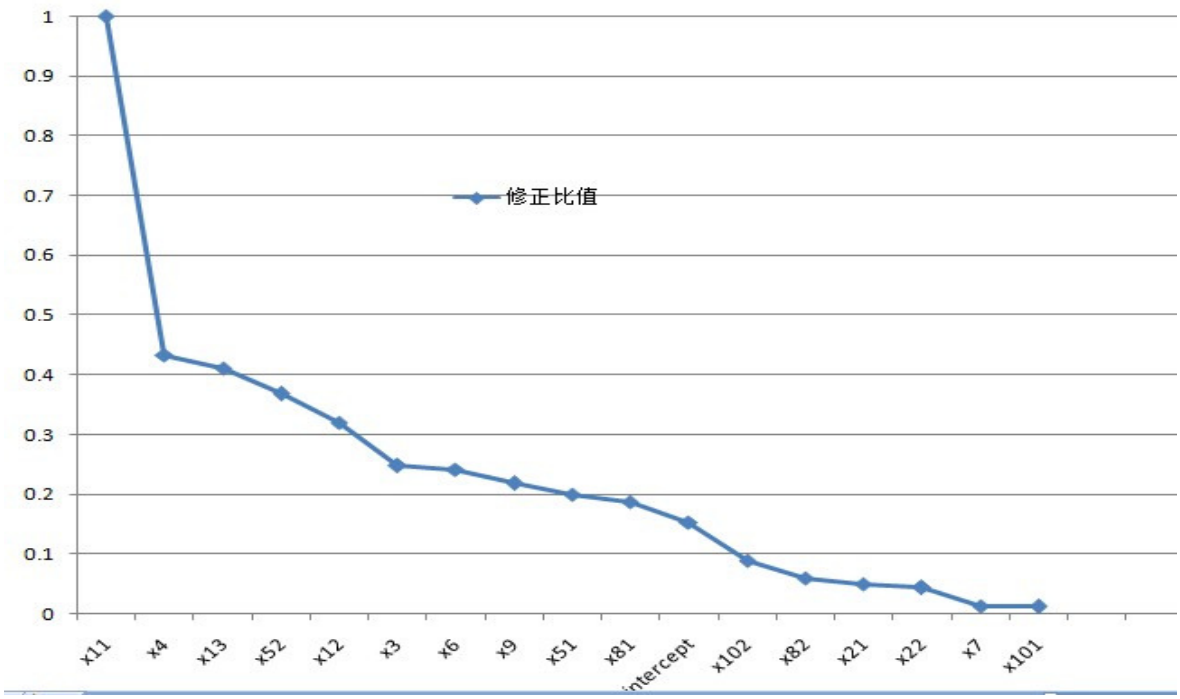
有了以上的理论基础，我们采用 R 统计软件使用逻辑回归进行编程，运行得到各参数估计如下：

表 11 各参数估计及检验结果

因素名称	参数估计	标准差	参数值/标准差的绝对值	修正比值
x11	-1.40E+02	1.33E+06	0.000105097	1
x4	-1.62E+01	3.57E+05	4.54316E-05	0.432280838
x13	1.05E+02	2.44E+06	4.31678E-05	0.410740972
x52	-8.79E+01	2.27E+06	3.86588E-05	0.367837189
x12	2.29E+02	6.82E+06	3.35583E-05	0.319306765
x3	2.17E+01	8.35E+05	2.60156E-05	0.24753767
x6	3.37E+01	1.33E+06	2.53534E-05	0.241236901
x9	-2.07E+01	9.04E+05	2.28786E-05	0.217689742
x51	-7.71E+01	3.69E+06	2.08864E-05	0.198733831
x81	-3.90E+01	1.98E+06	1.96472E-05	0.186942473
intercept	-1.63E+02	1.01E+07	1.60809E-05	0.153009873
x102	3.09E+01	3.30E+06	9.38126E-06	0.089262447
x82	1.21E+01	1.93E+06	0.00000625	0.059468616
x21	3.33E+01	6.36E+06	5.23038E-06	0.049766974
x22	-3.95E+00	8.57E+05	4.60829E-06	0.043847776
x7	-3.16E+00	2.45E+06	1.28618E-06	0.012237977
x101	2.89E+00	2.43E+06	1.18871E-06	0.011310553

此表中第四列为参数估计值和标准差的比值的绝对值，这个值越大意味着参数取值为非 0 的可能性就越大，从某种程度上说是变量重要程度的一种度量，为了更清晰地对这个比值大大小进行观察，我们给出它们相对于最大值的一个修正比值，并将其在图表中展示出来。此表是按照修正比值从大到小进行了排列，从这个排列我们可以得出：食道静脉曲张、Anti-HCV 以及 HbsAg 是否阳性、肿瘤位置、肿瘤大小、肿瘤的包膜是否完整等都对分类器具有很大的影响。其中若变量前面的系数为正数，说明具有该特征具有更大可能性的预后影响，相反地，变量系数为负说明具有该特征能够降低预后的影响。

我们拿食道静脉曲张来说，x11 是负值，x12 和 x13 都是正值，说明没有食道静脉曲张的患者比有（哪怕是轻微程度）曲张的患者预后效果会更好，也就是说预后没有影响的可能性更大。x51 和 x52（整个肝脏的系数是 0）的值都是负值，这说明单单只有左肝或者右肝的患者相对于整个肝部都有肿瘤的患者预后效果会更好。Anti-HCV(x4)为阳性，HbsAg 为阴性的患者会具有更好的预后效果。其次，数据说明肿瘤越大对预后影响也越大。



图二 修正比值图

从图二可以看出，这里的变量的重要程度相差很大，再加上各变量之间可能存在相关性，为了让模型更加优化，我们考虑逐步回归方法重新进行建立统计模型。运行程序后有以下结果：

表 12 各参数估计及检验结果

因素名称	参数估计	标准误差	比值
截距项	-185.47	246918.45	0.000751
x11	-197.19	275788.67	0.000715
x12	241.82	302195.18	0.0008
x3	94.56	153322.28	0.000617
x4	-93.47	128592.47	0.000727
x6	40.59	53090.49	0.000765
x82	46.51	91038.89	0.000511
x9	-142.80	189956.22	0.000752

对比优化前的系数的比值（参数值/标准差），可见优化后的系数更有可能不为 0，从而优化的模型会相比优化前的模型更好。同时我们也可知:食道静脉曲张、HbsAg 和 Anti-HCV 的阴阳性、肿瘤大小、肿瘤的包膜否完整以及肿瘤旁的微小子灶等都对预后具有很大影响。

为此，我们得到两个模型如下，分别为全模型（没有优化的模型）和逐步回归简化模型

$$p(Y^* = 1 \mid X_i^*) \approx p(X_i^{*\prime} \hat{\beta}) = \frac{e^{X_i^{*\prime} \hat{\beta}}}{1+e^{X_i^{*\prime} \hat{\beta}}}$$

且$\hat{\beta}$即为系数向量，其值分别由上述的表中的参数估计一列值给出，如对于优化的模型有：

$$X_i^{*\prime}\hat{\beta}=-185.47-197.19x11+241.82x12+94.56x3-93.47x4+40.59x6+46.51x82-142.8x9$$

2.3.2 预测评估

实际工作中，逻辑回归的一个很重要应用就是预测。在这个案例中，我们能否利用已经建立的逻辑回归对预后影响进行预测？如果能，那么可以为患者根据自己的实际情况决定最佳的治疗方案。要达到此目的，必须具备两种能力：第一，具备预测能力；第二，具备对预测精度评估的能力。

由于只有 20 个观测，假如我们直接把这 20 个作为训练样本又作为测试样本，这既当运动员又当裁判的做法不符合我们严谨的态度，假如 8:2 来分开训练样本和测试样本，这也会导致原本就少的观测变得更少，而且测试样本也过少。为此，我们考虑留一法进行建模与预测，即每个观测都充当一回测试样本，其余的进行建模，这样 20 次下来，我们得到两种模型对应的预测结果，即$p(Y^*=1|X_i^*)$，其中 yhat_all 为全模型的预测结果，yhat_step 为简化优化模型的结果，y 为最终的判定值 1 为有影响，0 为无影响。

表 13 两种模型留一检验对应的预测结果

No.	y	yhat_all	yhat_step	No.	y	yhat_all	yhat_step
1	1	2.220446e-16	4.449244e-11	11	0	1.644294e-12	1.000000e+00
2	0	2.220446e-16	2.220446e-16	12	1	1.000000e+00	1.000000e+00
3	1	1.000000e+00	1.000000e+00	13	0	7.117576e-09	9.917413e-11
4	1	5.259190e-03	7.371268e-12	14	0	1.000000e+00	2.220446e-16
5	0	4.682617e-03	9.078459e-10	15	1	2.220446e-16	4.194752e-07
6	1	1.000000e+00	1.000000e+00	16	0	1.000000e+00	2.220446e-16
7	1	1.000000e+00	1.000000e+00	17	0	1.000000e+00	1.000000e+00
8	0	2.220446e-16	2.220446e-16	18	1	2.220446e-16	5.443514e-11
9	0	1.000000e+00	2.220446e-16	19	0	1.501922e-01	2.220446e-16
10	0	2.220446e-16	9.917413e-11	20	1	2.220446e-16	1.000000e+00

此概率$p(Y^*=1|X_i^*)$量化了肝癌患者治疗预后影响的可能性。显然，如果该可能性很大，我们更倾向于将Y_i^*预测为$\hat{Y_i^*}=1$；否则将Y_i^*预测为$\hat{Y_i^*}=0$，但是到底多大的概率才叫大呢？为此，我们需要设置一个阈值α，再定义一个预测规则如下：

$$\hat{Y_i^*}=\begin{cases}1,\ p(X_i^{*\prime}\hat{\beta})>\alpha\\ 0,\ p(X_i^{*\prime}\hat{\beta})\le\alpha\end{cases}$$

如何选取阈值α呢？显然，我们的目标是预测的准确。那么如何评判预测结果的准确性呢？不同的评判标准，量化手段会产生不同阈值α.

1. 基于概率阈值α的错判率(mis-classification rate,MCR)

$$MCR=\frac{1}{m}I\left(\hat{Y}_i^*\neq Y_i^*\right)$$

其中m为预测样本总数，$I\left(\hat{Y}_i^*\neq Y_i^*\right)$为示性函数。

很明显，MCR是错误判断的比率。如果MCR=0，意味着所有的预测都正确；MCR=1，意味着所有预测都错误。我们的目标就是要极小化MCR，找到最优的阈值α.我们可以从0到1去搜索最优的阈值α，使得MCR极小化即可，通过编程可得阈值范围（0，1）。

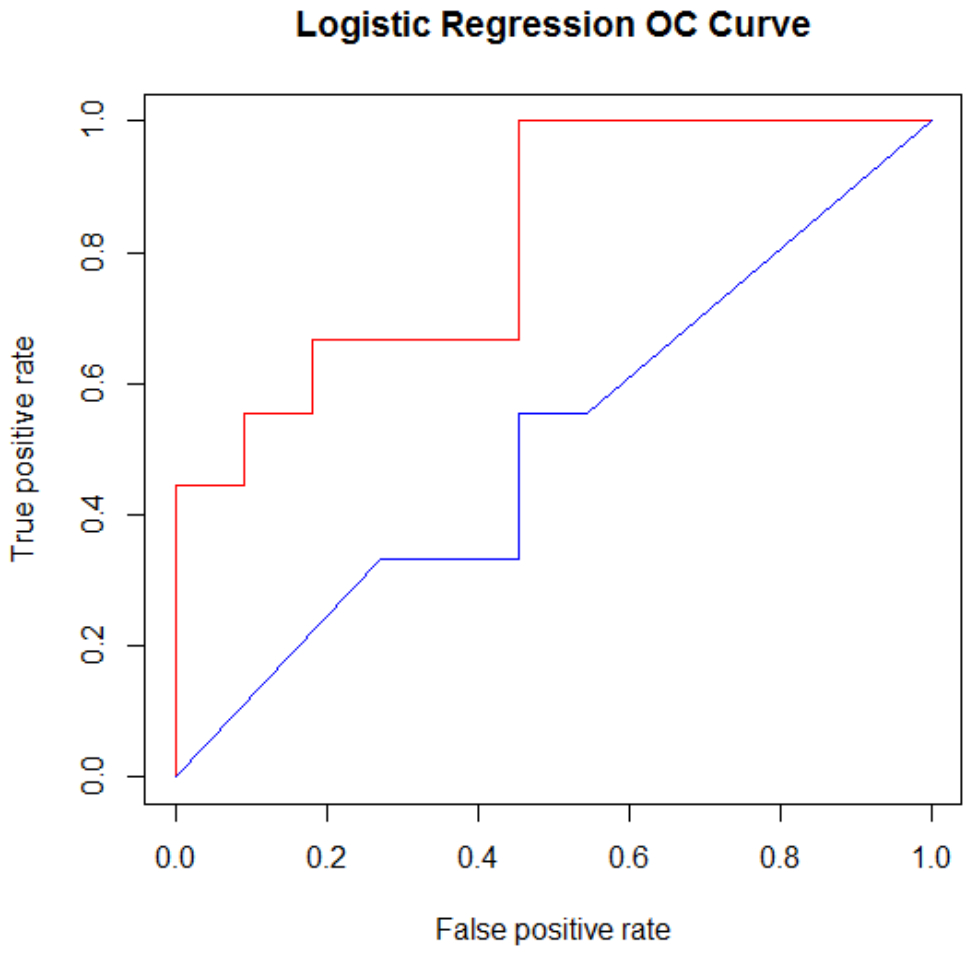
2. 基于接受者曲线（ROC）模型效果评判

虽然我们可以按照上面所说的方法可以找到阈值α最优解，但是我们这样考虑，如果建立的模型是这样的话，那么我们意味着把有影响划分到没有影响那类以及把没有影响划分到有影响那类是一样看待的。事实上，我们当然希望能够把对预后真实有影响的患者划分到有影响那一类，这样以尽早发现这类患者，才好做好各方面的医疗，降低预后的影响。而对于事实上没有影响的患者，我们也不希望把他划到有影响的一类，这样白白浪费医疗资源，而且会给患者带来沉重的经济负担。用数学语言来表达的话，即我们关注真实的$Y_i^*=0$预测为$\hat{Y}_i^*$=1 的概率$p(\hat{Y}_i^*=1|Y_i^*=0)$以及真实的$Y_i^*=1$预测为$\hat{Y}_i^*$=1的概率$p(\hat{Y}_i^*=1|Y_i^*=1)$。我们自然希望TPR（True positive rate）$=p(\hat{Y}_i^*=1|Y_i^*=1)$的概率越大越好，而FPR（False positive rate）$=p(\hat{Y}_i^*=1|Y_i^*=0)$越小越好。对于任给的一个阈值$\alpha$，我们相应地可以算出TPR以及FPR的值，于是把（FPR，TPR）在坐标轴上展示出来，越靠近左上方的点其效果越好，此时ROC曲线下面积也就是越大说明分类效果越好。为此，我们画出优化模型留一法得出的ROC曲线图，如图三：

图中红色曲线是优化后模型采用留一法得到的ROC曲线，而蓝色曲线是没有优化的模型留一法画出的ROC曲线。其中红色ROC曲线下面积为0.7802，而蓝色曲线下面积为0.6538，从这两个数据来看并结合图表来看，用筛选后的这几个变量来建立逻辑回归的分类器，具有一定的准确性，效果还是比较明显的，有一定的可泛化性。而全模型ROC曲线下的面积说明分类器的准确性较差，说明优化模型相对更好。

到此为止，我们已经确定选择优化模型作为我们较好的预测模型，但阈值该选什么呢？在阈值的选择方面，我们按照（FPR，TPR）尽量往（0,1）靠近的原则，为此编程得到当阈值α从0.005到0.995变化时，留一法检验时，分类器都将9个有预后影响的预测为有影响的为5个预测为没有影响有四个，11个没有影响的预测为有影响的有2个没有影响的有9个，总体的预测正确率为70%,此时MCR=30%， 我们在这里取阈值α使得效果最好的取值的平均数，即取阈值α=0.5，这也符合

我们的常识，即当预测为 1 类的概率大于 50%时，我们就将该患者预测为 1 类，否则预测为 0 类。



图三 ROC 曲线

3. 结论

3.1. 模型结论

通过对结果的对比分析，我们最终得到如下的逻辑回归模型：

$$p(Y^* = 1 \mid X_i^*) \approx p(X_i^{*\prime}\hat{\beta}) = \frac{e^{X_i^{*\prime}\hat{\beta}}}{1+e^{X_i^{*\prime}\hat{\beta}}}$$

其中

$X_i^{*\prime}\hat{\beta} = -185.47 - 197.19x11 + 241.82x12 + 94.56x3 - 93.47x4 + 40.59x6 + 46.51x82 - 142.8x9$ 。

且分类器，即预测的分类规则为

$$\hat{Y}_i^* = \begin{cases} 1, & p(X_i^{*\prime}\hat{\beta}) > 0.5 \\ 0, & p(X_i^{*\prime}\hat{\beta}) \le 0.5 \end{cases}$$

假如全部样本都拿来做训练样本并用原样本做检验，可得全部的样本都预测正确，即：

$$MCR = \frac{1}{m} I\left(\hat{Y}_i^* \neq Y_i^*\right) = 0。$$

该分类器交叉验证（留一检验）错判率为：

$$MCR = \frac{1}{m} I\left(\hat{Y}_i^* \neq Y_i^*\right) = 0.3，$$

其 ROC 曲线图如上图，ROC 曲线下面积为 0.7802 可见这个分类器效果还是比较准确的，具有比较优良的泛化性。

以上的结论表明，这模型的预测效果还是不错的。若希望进一步提高模型的应用，我们可以通过以逻辑回归判别模型为基模型，进行深入的 bagging 方法进行建模，产生更强的分类器，以帮助分析，得出更有建设性的结论。

3.2. 手术和治疗方案建议

通过逻辑回归法的建模分析，我们发现是否有食道静脉曲张、HbsAg 和 Anti-HCV 的阴阳性、肿瘤大小、肿瘤的包膜否完整以及肿瘤旁的微小子灶等都对预后具有很大的影响。因此，我们提出如下的手术和治疗方案建议：

1. HbsAg 和 Angi-HCV 分别为为乙肝表面抗原和丙肝表面抗原，如果阳性代表有乙肝病毒和丙肝病毒。而感染乙肝和丙肝病毒对肝细胞伤害大，需要重点进行抗病毒处理。如果已经出现肝癌，需要进行手术，医生应注意手术期间的防护以免发生感染。
2. 是否有肿瘤旁的微小子灶，肿瘤的包膜是否完整这三个指标都是判断癌细胞是否发生转移的指标，其中肿瘤旁的微小子灶是肝癌发生肝内转移的地方。如果癌细胞发生转移，那么治疗方案中需要在原来的基础上补充进行化疗。而且出现转移提示病情已经比较严重。
3. 存在食道静脉曲张是肝硬化和肝癌的一个临床表现，表明干细胞功能减退，引起水钠滞留，同时也提示病人病情比较严重，随时有上消化道出血并发症的可能。因此在手术后可以做手术改善处理，可以采取 TIPS、分流和断流三种不同的方式。

4. 参考文献

[1]. Alan Julian Izenman . Modern Multivariate Statisitical Techniques,Philadelphia,Pennsylvabia,April 2008.

5. 附录

5.1 附录一

#描述统计 in R #

```
library(sqldf)
a=read.table("t3.txt",header=T)
a$y=(a$DECISION=="Y")
sqldf(" select X1,count(*), avg(y) from a group by X1 order by avg(y) ")
sqldf(" select X2, count(*),avg(y) from a group by X2 order by avg(y) ")
sqldf(" select X3, count(*),avg(y)  from a group by X3 order by avg(y) ")
sqldf(" select X4,count(*), avg(y) from a group by X4 order by avg(y) ")
sqldf(" select X5, count(*),avg(y) from a group by X5 order by avg(y) ")
sqldf(" select X6,count(*), avg(y) from a group by X6 order by avg(y) ")
sqldf(" select X7, count(*),avg(y) from a group by X7 order by avg(y) ")
sqldf(" select X9, count(*),avg(y) from a group by X9 order by avg(y) ")
sqldf(" select X8, count(*),avg(y) from a group by X8 order by avg(y) ")
sqldf(" select X10,count(*), avg(y) from a group by X10 order by avg(y) ")
```

5.2 附录二

```
library(rpart)
library(MASS)
library(ROCR)
library(pROC)
library(sqldf)
#
a=read.table("t3.txt",header=T)
a$x11=1*(a$X1=='no')
a$x12=1*(a$X1=='light')
a$x13=1*(a$X1=='mid')
a$x21=1*(a$X2=='no')
a$x22=1*(a$X2=='branch')
a$x3=1*(a$X3=='positive')
a$x4=1*(a$X4=='positive')
a$x51=1*(a$X5=='leftliver')
a$x52=1*(a$X5=='rightliver')
a$x6=1.5*(a$X6=='small')+4*(a$X6=='middle')+7.5*(a$X6=='big')+10*(a$X6=='verybig')
a$x7=1*(a$X7=='dilation')
a$x81=1*(a$X8=='integrate')
a$x82=1*(a$X8=='part')
a$x9=1*(a$X9=='have')
a$x101=1*(a$X10=='no')
a$x102=1*(a$X10=='less')
a$y=1*(a$DECISION=='Y')

dim(a)
data=a[,c(12:28)]
dim(data)

#logistic regression for all valueables
mylogit <- glm(y~., data = data, family = binomial(logit))
summary(mylogit)

mylogit1 <- step(glm(y~., data = data, family = binomial(logit)))
summary(mylogit1)
```

```
fmla <-paste(names(data)[17],"~", paste(names(data)[1: length(names(data[,1:17]))-1], collapse="+"))
fmla <- as.formula(fmla)

#留一交叉检验
for(i in 1:20){
#split function to split dataset into training set and validate set
splitdf <- function(dataframe, i) {
  validateindex=i
  trainset <- dataframe[-validateindex, ]
  validateset <- dataframe[validateindex, ]
  list(trainset=trainset,validateset=validateset)
}
splitdata <- splitdf(data, i)
train_data <- splitdata$trainset
validate_data <- splitdata$validateset

#logistic regression
mylogit <- glm(y~x11+x12+x3+x4+x6+x82+x9, data = train_data, family = binomial(logit))

predict_mylogit <- predict(mylogit, newdata = validate_data, type = "response")
if (i==1) {
predict_mylogit1=predict_mylogit  }
if (i==2) {
predict_mylogit2=predict_mylogit  }
if (i==3) {
predict_mylogit3=predict_mylogit  }
if (i==4) {
predict_mylogit4=predict_mylogit  }
if (i==5) {
predict_mylogit5=predict_mylogit  }
if (i==6) {
predict_mylogit6=predict_mylogit  }
if (i==7) {
predict_mylogit7=predict_mylogit }
if (i==8) {
predict_mylogit8=predict_mylogit }
if (i==9) {
predict_mylogit9=predict_mylogit  }
if (i==10) {
predict_mylogit10=predict_mylogit  }
if (i==11) {
predict_mylogit11=predict_mylogit  }
if (i==12) {
predict_mylogit12=predict_mylogit  }
if (i==13) {
predict_mylogit13=predict_mylogit  }
if (i==14) {
predict_mylogit14=predict_mylogit  }
if (i==15) {
predict_mylogit15=predict_mylogit  }
if (i==16) {
predict_mylogit16=predict_mylogit  }
if (i==17) {
predict_mylogit17=predict_mylogit  }
if (i==18) {
predict_mylogit18=predict_mylogit  }
if (i==19) {
predict_mylogit19=predict_mylogit  }
if (i==20) {
```

```
predict_mylogit20=predict_mylogit  }
}

predict_mylogit0=c(predict_mylogit1,predict_mylogit2,predict_mylogit3,predict_mylogit4,predict_mylog
it5,predict_mylogit6,predict_mylogit7,predict_mylogit8,predict_mylogit9,predict_mylogit10,predict_my
logit11,predict_mylogit12,predict_mylogit13,predict_mylogit14,predict_mylogit15,predict_mylogit16,pr
edict_mylogit17,predict_mylogit18,predict_mylogit19,predict_mylogit20)

for(i in 1:20){
#split function to split dataset into training set and validate set
splitdf <- function(dataframe, i) {
  validateindex=i
  trainset <- dataframe[-validateindex, ]
  validateset <- dataframe[validateindex, ]
  list(trainset=trainset,validateset=validateset)
}
splitdata <- splitdf(data, i)
train_data <- splitdata$trainset
validate_data <- splitdata$validateset

#logistic regression
mylogit <- glm(fmla, data = train_data, family = binomial(logit))
predict_mylogit <- predict(mylogit, newdata = validate_data, type = "response")
if (i==1) {
predict_mylogit1=predict_mylogit  }
if (i==2) {
predict_mylogit2=predict_mylogit  }
if (i==3) {
predict_mylogit3=predict_mylogit  }
if (i==4) {
predict_mylogit4=predict_mylogit  }
if (i==5) {
predict_mylogit5=predict_mylogit  }
if (i==6) {
predict_mylogit6=predict_mylogit  }
if (i==7) {
predict_mylogit7=predict_mylogit }
if (i==8) {
predict_mylogit8=predict_mylogit }
if (i==9) {
predict_mylogit9=predict_mylogit  }
if (i==10) {
predict_mylogit10=predict_mylogit  }
if (i==11) {
predict_mylogit11=predict_mylogit  }
if (i==12) {
predict_mylogit12=predict_mylogit  }
if (i==13) {
predict_mylogit13=predict_mylogit  }
if (i==14) {
predict_mylogit14=predict_mylogit  }
if (i==15) {
predict_mylogit15=predict_mylogit  }
if (i==16) {
predict_mylogit16=predict_mylogit  }
if (i==17) {
predict_mylogit17=predict_mylogit  }
if (i==18) {
predict_mylogit18=predict_mylogit  }
```

```
if (i==19) {
predict_mylogit19=predict_mylogit  }
if (i==20) {
predict_mylogit20=predict_mylogit  }
}
predict_mylogit1=c(predict_mylogit1,predict_mylogit2,predict_mylogit3,predict_mylogit4,predict_mylogit5,predict_mylogit6,predict_mylogit7,predict_mylogit8,predict_mylogit9,predict_mylogit10,predict_mylogit11,predict_mylogit12,predict_mylogit13,predict_mylogit14,predict_mylogit15,predict_mylogit16,predict_mylogit17,predict_mylogit18,predict_mylogit19,predict_mylogit20)

cbind(predict_mylogit1,predict_mylogit0)

#ROC
#ROC
pred<- prediction(predict_mylogit0,data[,17])
perf<- performance(pred, "tpr","fpr" )
pred1<- prediction(predict_mylogit1,data[,17])
perf1<- performance(pred1, "tpr","fpr" )
plot(perf)
plot(perf,  main="Logistic  Regression  OC  Curve",xlab="False  positive  rate",  ylab="True  positive rate",col="red")
plot(perf1,col="blue",add=T)
roc(data[,12], predict_mylogit0)
roc(data[,12], predict_mylogit1)



for(i in 1:20){
print(i*0.05)
print(table((predict_mylogit0>0.05*i)*1,data$y))
}
```