卡方检验的计算

[卡方检验\_c语言实现卡方检验-CSDN博客](https://blog.csdn.net/qq_37746942/article/details/115537596?ops_request_misc=%7B%22request%5Fid%22%3A%22170541548716800222827315%22%2C%22scm%22%3A%2220140713.130102334.pc%5Fall.%22%7D&request_id=170541548716800222827315&biz_id=0&utm_medium=distribute.pc_search_result.none-task-blog-2~all~first_rank_ecpm_v1~rank_v31_ecpm-1-115537596-null-null.142^v99^pc_search_result_base4&utm_term=%E5%8D%A1%E6%96%B9%E6%A3%80%E9%AA%8Cc++&spm=1018.2226.3001.4187)

p值的理解

如何理解P值的定义? - Patrick的回答 - 知乎

<https://www.zhihu.com/question/405665370/answer/2428242071>

问题

对于gender变量，原始的"M"和"F"被转换成了factor类型的0和1，然后通过数值标签表示男性和女性。  
后续，factor类型的gender通过as.numeric转换为数值变量，变成了0和1，但此时R不再知道0对应于男性还是女性，只是简单地将它们看作数值0和1。

**双向表卡方检验**（Chi-squared Test of Independence）用于检验两个分类变量之间是否存在关联或独立。具体而言，它检查两个分类变量之间的观察值分布是否与期望的分布有显著的差异。  
问题  
在双向表卡方检验中，我们考虑两个分类变量，并建立一个观察频数表（双向表）来记录每个组合的观察频数。然后，我们计算期望的频数，这是在两个变量独立的情况下，每个单元格的期望频数。最后，通过比较观察频数和期望频数，使用卡方统计量来判断两个变量之间是否存在显著关联。  
  
具体步骤如下：  
  
建立双向表（观察频数表）： 将两个分类变量的所有可能组合列成一个表格，记录每个组合的观察频数。  
  
计算期望频数： 计算在两个变量独立的假设下，每个单元格的期望频数。期望频数的计算通常基于总体的边际分布。  
  
计算卡方统计量： 使用观察频数和期望频数计算卡方统计量。卡方统计量表示观察频数与期望频数之间的偏离程度。  
  
进行假设检验： 利用卡方统计量进行假设检验。假设检验的零假设是两个变量独立。根据卡方分布，计算 p-value，若 p-value 小于显著性水平（通常为0.05），则拒绝零假设，认为两个变量之间存在显著关联。

问题

# 将分类变量转化为数值变量  
data2$status = as.numeric(data2$status) - 1  
data2$ssc\_b = as.numeric(data2$ssc\_b) - 1  
data2$hsc\_b = as.numeric(data2$hsc\_b) - 1  
data2$hsc\_s = as.numeric(data2$hsc\_s) - 1  
data2$degree\_t = as.numeric(data2$degree\_t) - 1  
data2$specialisation = as.numeric(data2$specialisation) - 1  
data2$gender = as.numeric(data2$gender) - 1  
data2$workex = as.numeric(data2$workex) - 1

在这种情况下，它们的映射值是连续的整数，但由于这是一个分类变量，这种映射可能给模型带来误导。如果模型以为这是一个有序的连续变量，可能会错误地给予不同水平之间的值大小关系。通常情况下，分类变量的编码不应该给予任何大小关系，而只是表示不同的类别。因此，通过减一将其映射为 0、1、2，更符合分类变量的本质。

举个例子例如degree\_t Copy code  
data2$degree\_t = as.numeric(data2$degree\_t) - 1  
在这里，degree\_t 是一个包含三个水平的因子变量，它的水平是 "Comm&Mgmt"，"Others"，和 "Sci&Tech"。通过上述代码，进行了如下映射：  
"Comm&Mgmt" 被映射为 0  
"Others" 被映射为 1  
"Sci&Tech" 被映射为 2

问题

为什么选取的特征数为8要过滤掉5  
  
在 AUC（曲线下面积）值的图中，有一个虚线，通常是在横坐标上表示选取的特征数。描述中提到，在 λ 取第一条虚线处的值时，模型的 AUC 值最大。这意味着在某个正则化水平（由 λ 决定）下，模型的性能最优，对应于选取了特定数量的特征。在这个例子中，选取的特征数为8。