# 第三阶段 从数据分析到数据挖掘

# SPSS-[数据分析2]

## 网校相关视频

1. 相关分析
2. 回归分析
3. Logistics回归

## 明确本次课知识点，明确重点难点

### 【知识点目标】

* SPSS回归分析
* SPSS 差异性检验

### 【重点】

* SPSS回归分析

### 【难点】

* SPSS回归分析

## 复习巩固作业讲解

#### SPSS相关分析简介

#### SPSS相关分析前置条件

#### SPSS相关分析基本操作

#### SPSS相关分析结果解读

#### SPSS回归分析简介

* SPSS线性回归分析
* SPSS一元线性回归分析
  + 一元线性回归分析前置条件
  + 一元线性回归分析基本操作
  + 一元线性回归分析结果解读
  + 用一元线性回归分析进行预测
* SPSS自动线性建模
  + 基本操作
  + 结果解读
  + 预测

## 本次课程任务讲解

### 【知识点1】SPSS回归分析

#### SPSS多重线性回归

多元线性回归：多元线性回归的自变量X的数据类型是连续型变量。

多重线性回归：多重线性回归的自变量X的数据类型可能存在多种数据类型，例如性别等的离散型变量。

自变量超过1个，因变量只有1个的时候，这种线性回归称为多重线性回归

其回归方程为：Y=a+b1x1+b2x2+....+bnxn+e

* 1. 多重线性回归分析前置条件

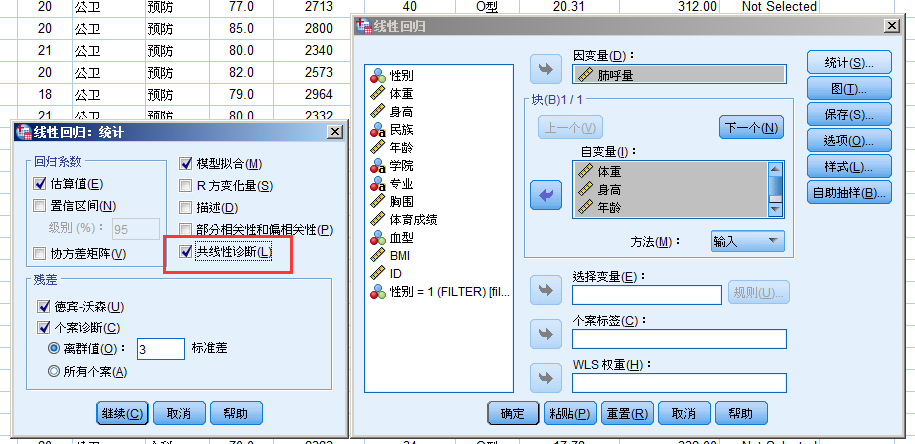
多重线性回归中允许自变量存在分类变量

其余与一元线性回归类似

* 1. 多重线性回归分析基本操作

和一元线性回归基本类似，但要注意增加多重共线性检查，另外要注意回归使用的方法



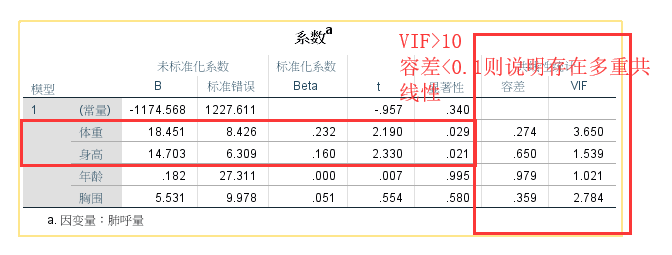


* 1. 多重线性回归分析结果解读



其他和一元线性回归类似

* 1. 多重线性回归分析中的多重共线性问题



如果存在比较严重的多重共线性问题，可以考虑将回归方法改为“步进”

* 1. 用多重线性回归分析进行预测

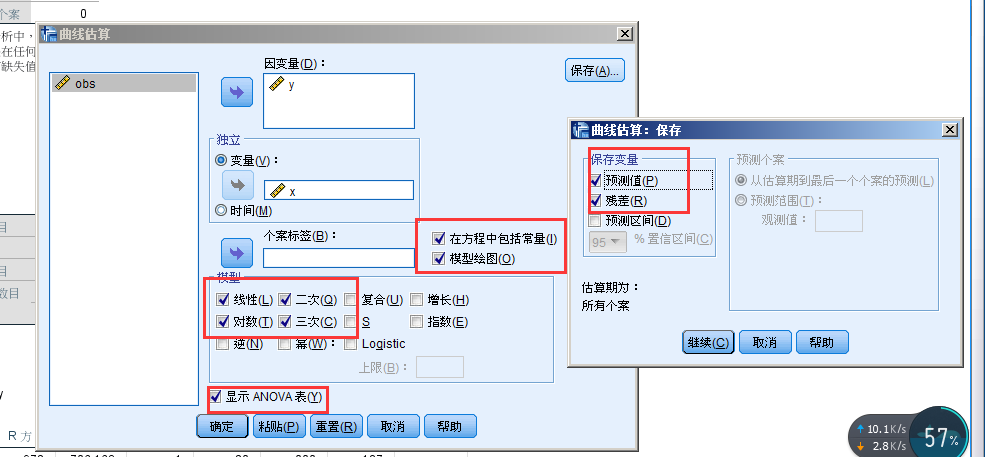
与一元线性回归类似

预测性回归分析：主要目的是为了通过给定新的自变量值来预测因变量的变化

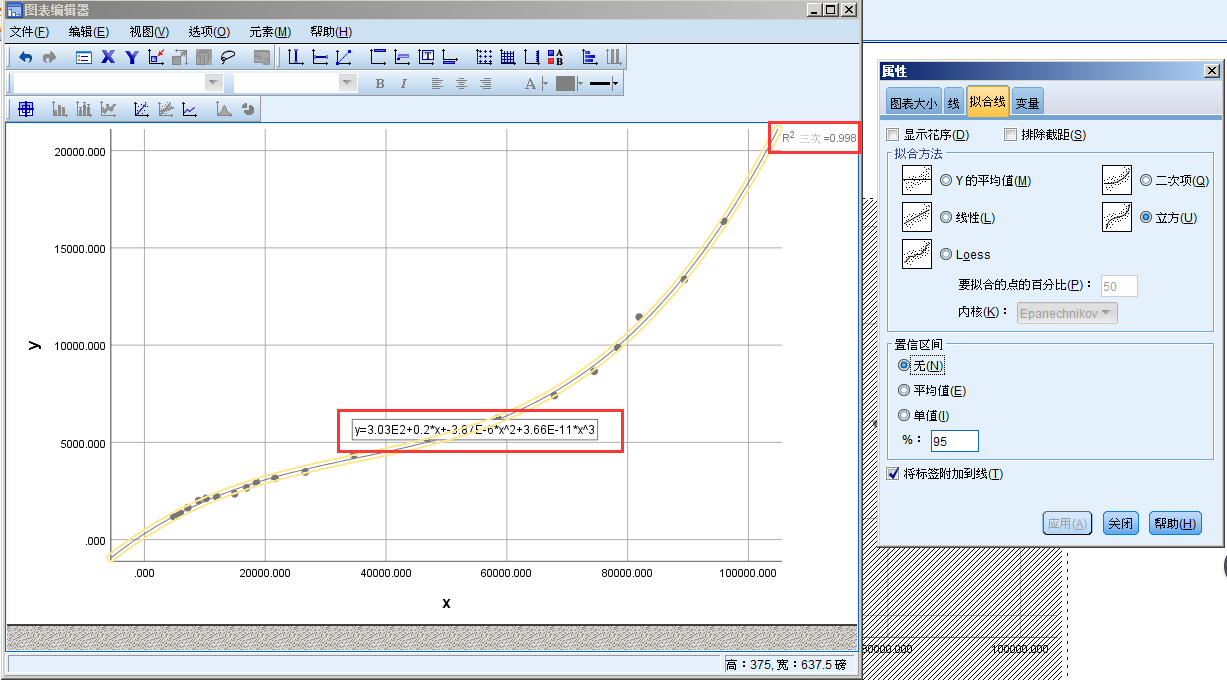
解释性回归分析：主要目的是为了观察哪些自变量对因变量会造成影响

1. SPSS曲线回归

用各种曲线方程拟合自变量与因变量之间的关系



通过R方或调整后R方判断拟合情况



1. SPSS 二元Logistic回归

因变量一定要是二分类的分类变量，自变量可以是任何变量

概率：P（Y/N）=0~1

不是用来进行预测，更多的是用来比较多个因素对因变量的影响情况

* 1. 二元Logistic回归分析前置条件

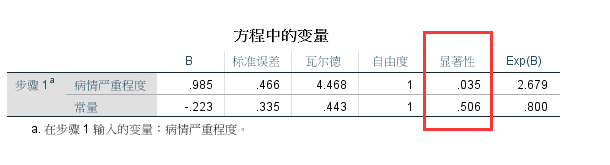
因变量的限制

* 1. 二元Logistic回归分析基本操作

先做单因素的Logistic回归分析，再考虑将哪些变量纳入回归方程

先看发病情况和病情严重程度的关系





病情严重程度的显著性为0.035<0.1(范围放大些),意味着这个变量对发病情况有显著性影响，所以应该放到回归方程中

同样的再去找x2/x3/x4对发病情况的影响

针对多分类的自变量，要拆分成多个二分类的哑变量，再纳入回归方程处理

X4(A,B,AB,O)

X4a(0,1)

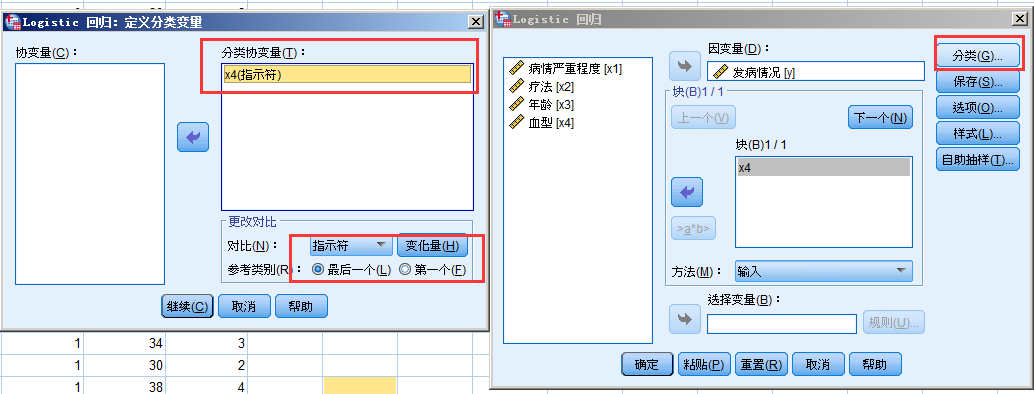
X4b(0,1)

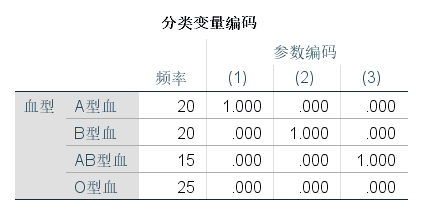
X4ab(0,1)

X4o(0,1)

一个N分类的自变量，需要创建N-1个二分类的自变量

（三个变量的值确定好后，第四个自动确定，一般默认第一个或者最后一个为基准）



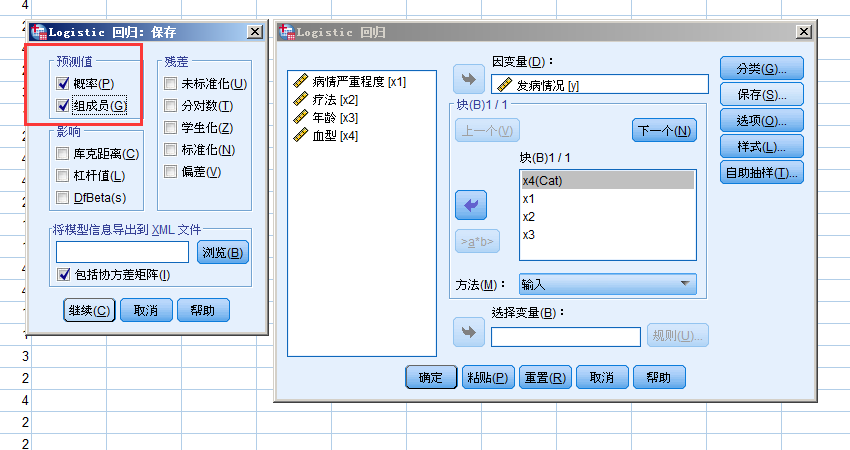


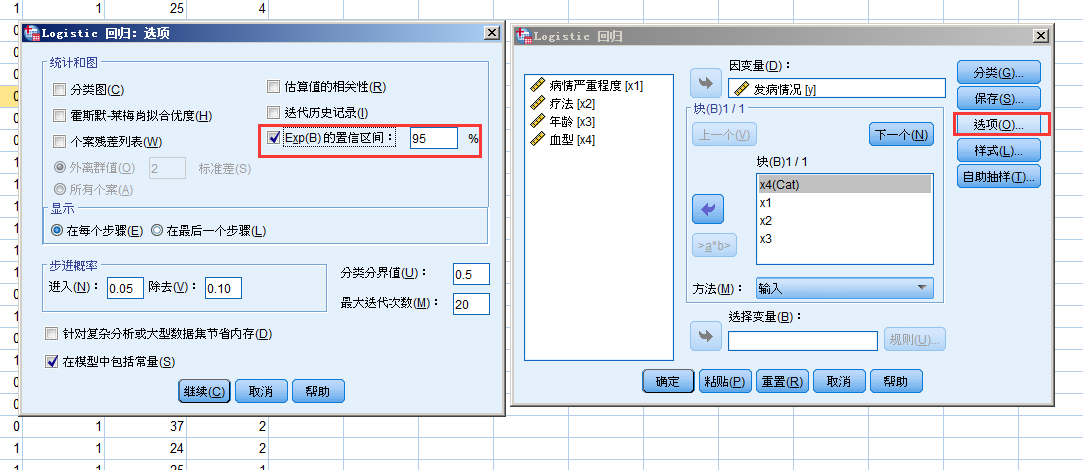
最后一个O型血作为参考，当A型血（1）、B型血（2）、C型血（3）都为0的时候，O型血为1



哑变量中只要有一个通过显著性检验（<0.1），通常会将其全部置入回归方程

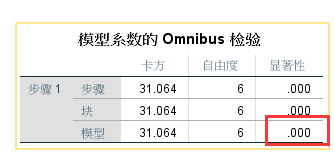
所有变量都放入进行观察





* 1. 二元Logistic回归分析结果解读

模型的显著性



模型显著性0.000<0.05,说明该模型有意义

模型的OR比：Exp(B)



Logistic回归方程意义不大，但OR比（说明自变量对因变量的发生风险）很重要。上述方程可解读为：

显著性都小于0.1，通过了显著性检验。

OR(病情程度)=3.864，意味着病情重的患者其病情复发风险是病情轻的患者的3.864倍

OR(疗法)=11.628，意味着使用传统疗法的患者病情复发风险是使用新疗法的患者的11.628倍

OR(血型(1))=2.559，意味着血型（1）（A型）相比O型血的复发风险是2.559倍。

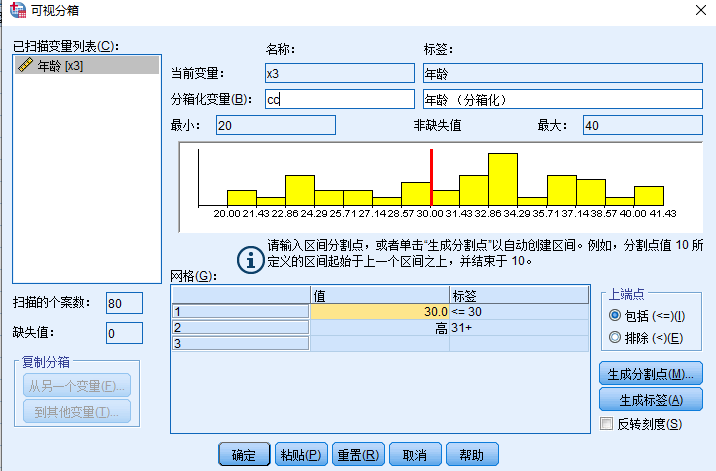
OR(血型(2))=7.3，意味着血型（2）（B型）相比O型血的复发风险是7.3倍。

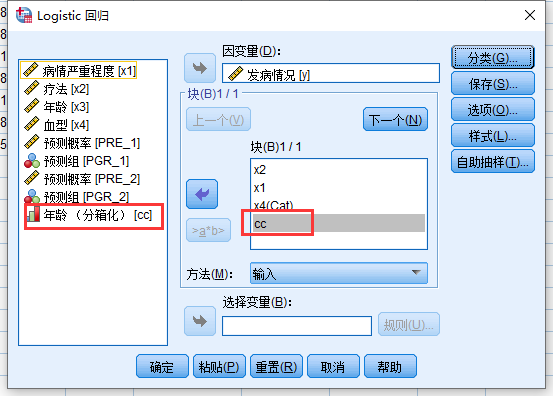
OR(血型(3))=8.282，意味着血型（3）（AB型）相比O型血的复发风险是8.282倍。

OR(年龄)=1.125，意味着年龄增加一岁，病情复发风险增加12.5%

（年龄一般很少研究增加一岁，增加的概率）

对年龄的处理方法：







年龄分箱，30岁以上是30岁以下病情复发的3倍多

## 自主学习作业讲解

无

## 课程总结

* SPSS多重线性回归分析
  + 多重线性回归分析前置条件
  + 多重线性回归分析基本操作
  + 多重线性回归分析结果解读
  + 用多重线性回归分析进预测
* SPSS曲线回归（拟合）
* SPSS Logistic回归
  + 二元Logistic回归前置条件
  + 二元Logistic回归基本操作
  + 二元Logistic回归结果解读

## 下次自主学习任务布置

#### 观看预习视频

第四、五、六章

#### 课后作业

见作业文档