# 第一阶段 业务分析师

# 建模数据分析师

## 网校相关视频

SPSS数据分析

## 明确本次课知识点，明确重点难点

### 【知识点目标】

* SPSS数据预分析
* 频率分析
* 描述分析
* 探索分析
* 交叉列联表分析
* 相关分析
* SPSS回归分析
* SPSS多元线性回归分析
* SPSS自动线性建模

### 【重点】

* SPSS数据预分析
* 频率分析
* 描述分析
* 探索分析
* 交叉列联表分析
* 相关分析
* SPSS回归分析
* SPSS多元线性回归分析
* SPSS自动线性建模

### 【难点】

* 相关分析
* SPSS回归分析
* SPSS多元线性回归分析

## 复习巩固作业讲解

无

## 本次课程任务讲解

### 【知识点1】SPSS数据预分析

**预分析是对数据进行深入分析的基础，可以通过简单的统计分析方法对数据的总体特征进行观察，以便选择更为深入的统计分析方法**

### 【知识点2】频率分析

针对分类变量的频率分析

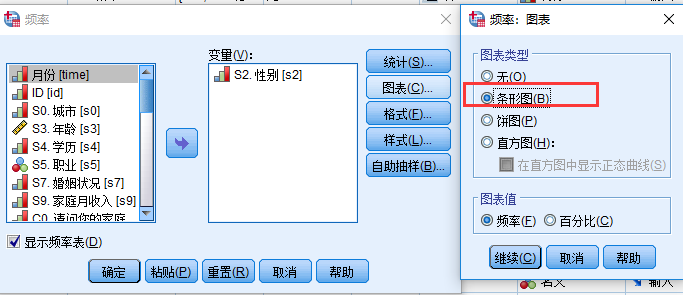
一般只有众数才有意义，多采用条形图或饼图进行观察

针对连续变量的频率分析

可以观察多个不同的统计量，可使用直方图观察数据的分布情况

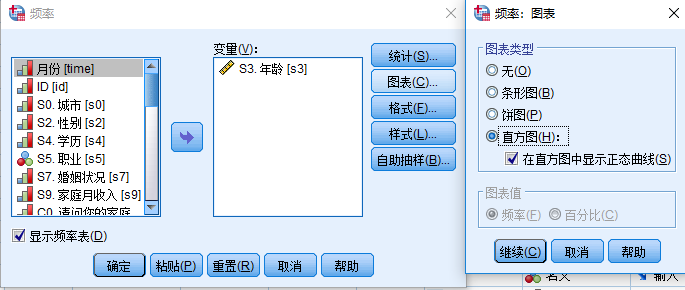
例子：查看CCSS\_Sample.sav文件性别的频率分析，以条形图的显示效果查看





例子：查看CCSS\_Sample.sav文件年龄的频率分析，以直方图的显示效果查看





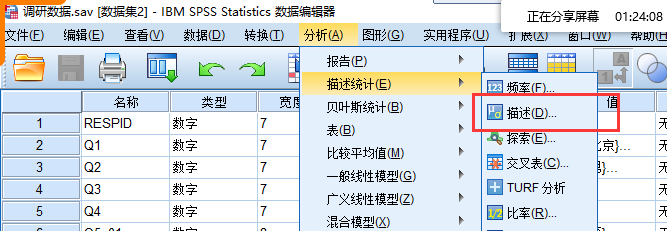
### 【知识点3】描述分析

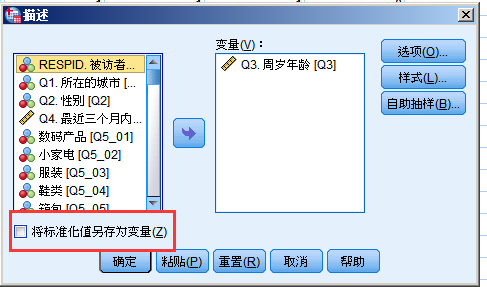
针对连续变量进行统计描述

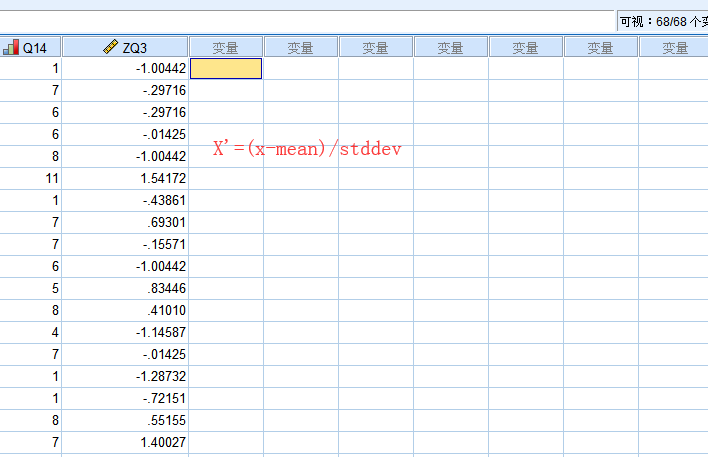
可以快速创建一个针对连续变量的Z标准化字段，对连续变量进行标准化处理

数据源：调研数据.sav

例子：描述周岁年龄



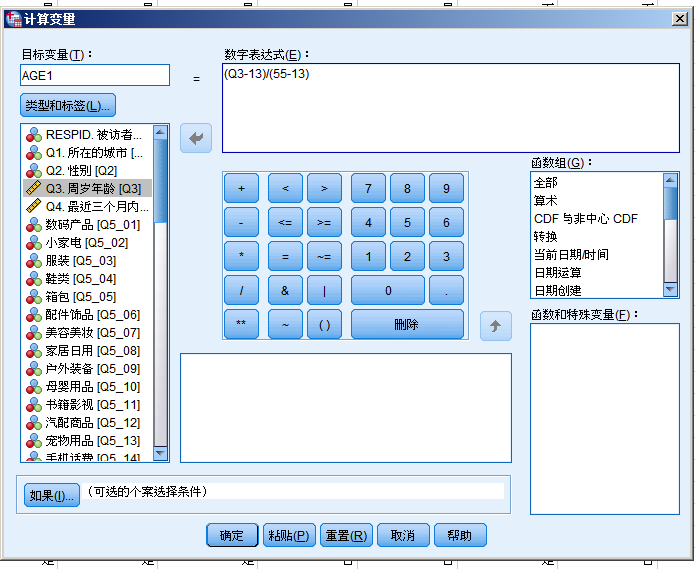




X’=(x-均值)/标准差

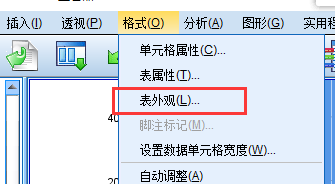
2、归一化：将连续变量映射到[0,1]区间

X’=(X-minX)/(maxX-minX)



表外观的设置：

1. 在输出窗口激活表格后，点击格式——表外观



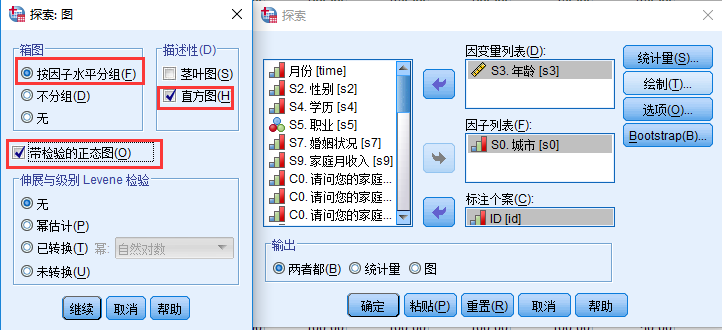
2）在数据视图的编辑——选项中设置成系统默认外观



### 【知识点4】探索分析

处理分组的箱图：分类变量对连续变量的影响（分组的连续变量分布情况）

范例：针对CCSS\_Sample.sav, 用探索过程考察不同城市受访者年龄的分布情况。



因变量列表：观察的变量——年龄

因子列表：分组的变量——城市

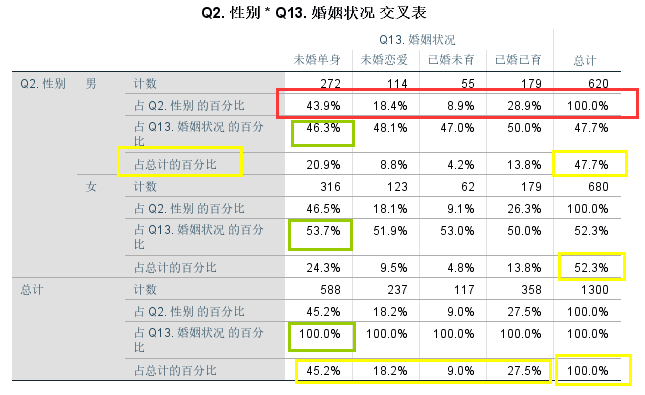
个案标注依据：id

### 【知识点5】交叉列联表分析

分析/观察两个或多个分类变量之间的关联关系（类似于EXCEL中的数据透视表）

范例：针对调研数据.sav, 用交叉表考察不同性别的婚姻情况。





### 【知识点6】相关分析

1、介绍：研究两个或两个以上随机变量之间的相互关系

线性相关/非线性相关

线性相关：研究两个连续变量之间的线性关系

2、SPSS相关分析前置条件

1. 两个变量都要是连续变量

可以直接确认

1. 两个连续变量是配对的，来自于同一个个体

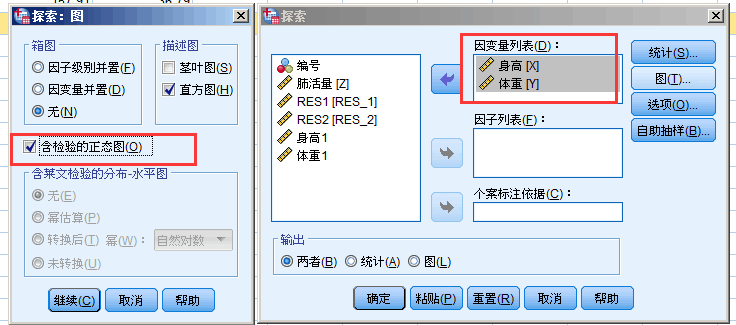
可以直接确认

1. 两个连续变量之间存在线性关系

散点图

1. 两个变量没有明显的异常值（相关系数容易受到异常值的影响）
2. 两个变量都要符合正态分布（探索分析）

素材：corr1.sav

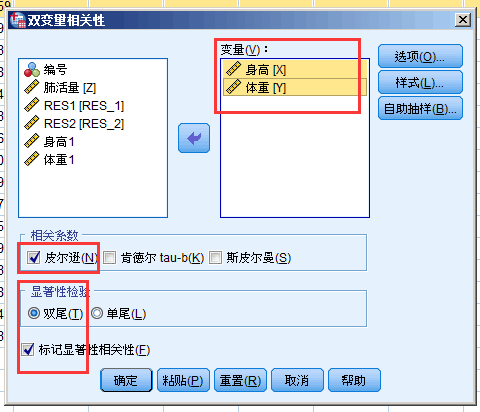


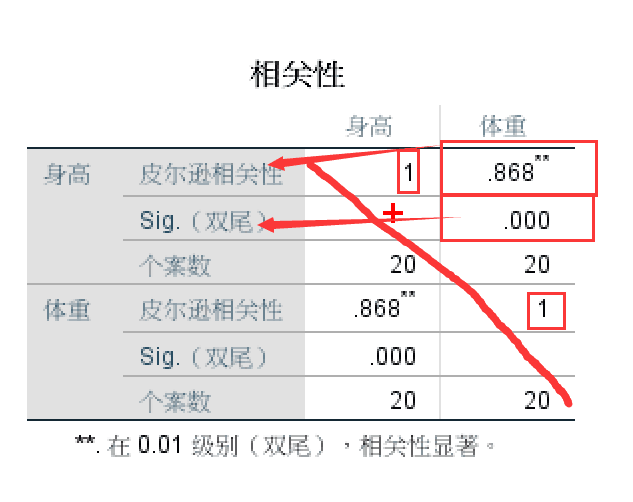


Shapiro-Wilk检验用于验证一个随机样本数据是否来自正态分布。

一般来说，如果数据接近正态分布，那么Shapiro-Wilk检验的P值就大于0.05；反之如果数据并不接近正态分布，那么Shapiro-Wilk检验的P值就小于0.05。

3、相关分析基本操作

位置：分析——相关——双变量



Pearson相关系数：取值在[-1,1],绝对值表示相关程度，0代表完全不相关，1代表完全相关，正负号表示相关方向。

在说明变量之间线性相关程度时，根据经验可将相关程度分为以下几种情况：

若|r|≥0.8 时，视为高度相关；若0.5≤|r|＜0.8 时，视为中度相关；

当0.3≤|r|＜0.5时，视为低度相关；当|r|＜0.3时，可视为不相关。

此外，在进行相关分析判定时，还需要关注Sig值（检验水准），该值表明相关判定的最大允许误差概率，通常取值为0.1/0.05/0.01，其判定方式与r类似（Sig>=0.1,0.1>Sig>=0.05,0.05>Sig>=0.01,Sig<0.01）。

sig值通常用 P>0.05 表示差异性不显著；0.01<P<0.05 表示差异性显著；P<0.01表示差异性极显著，具有统计学意义。

### 【知识点7】SPSS回归分析

#### SPSS回归分析简介

研究因变量对自变量的依赖关系，一般用来发现各种变量之间的因果关系或预测

1. SPSS线性回归分析

线性/非线性

一元/多重

一元线性回归研究的是一个连续变量（因变量）的取值随着另一个连续变量（自变量）的取值变化而产生线性变化的趋势

y=ax+b

1. SPSS一元线性回归分析
   1. 一元线性回归分析前置条件
      1. 因变量和自变量都是连续变量

直接确定

* + 1. 因变量和自变量之间存在线性关系

散点图观察

* + 1. 各观测值之间相互独立

独立性检验，在回归分析中处理

* + 1. 因变量没有显著的异常值

散点图观察

* + 1. 残差要符合正态分布且方差齐（预测值-实际值）

正态性检验，回归分析中处理

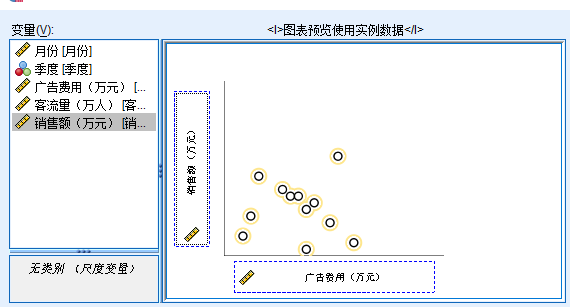
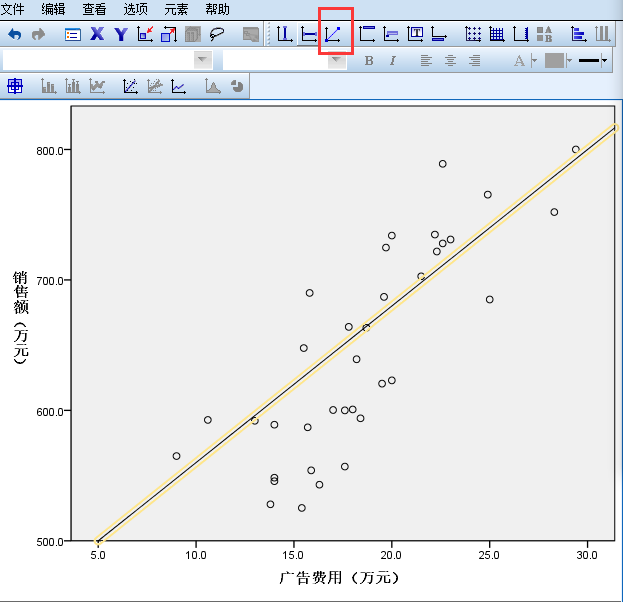
* 1. 一元线性回归分析基本操作

分析-回归-线性，数据：超市销售数据.sav

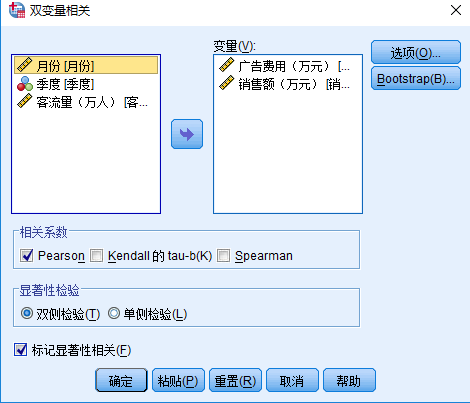
步骤：1）确定因变量和自变量  
自变量：广告费用

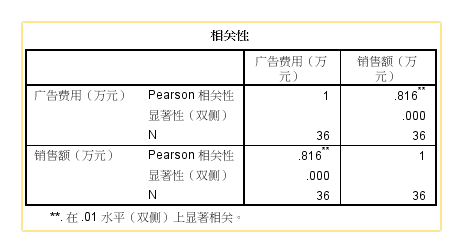
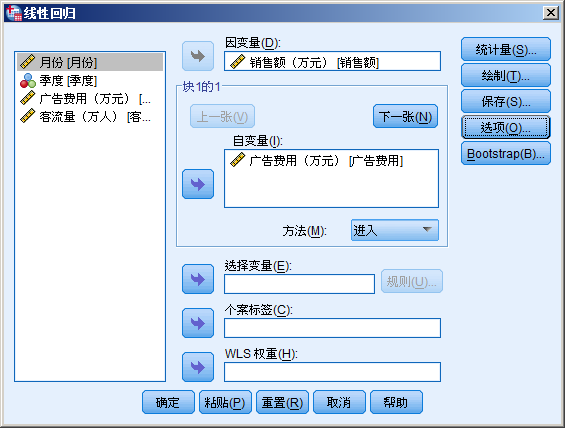
因变量：销售额

2）绘制散点图观察变量之间的关系

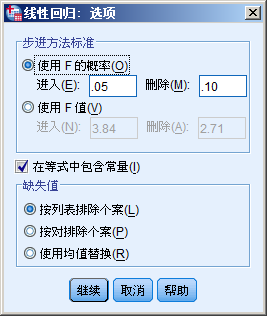
3）通过皮尔逊系数了解变量之间的相关关系

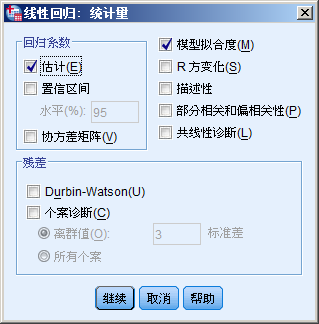




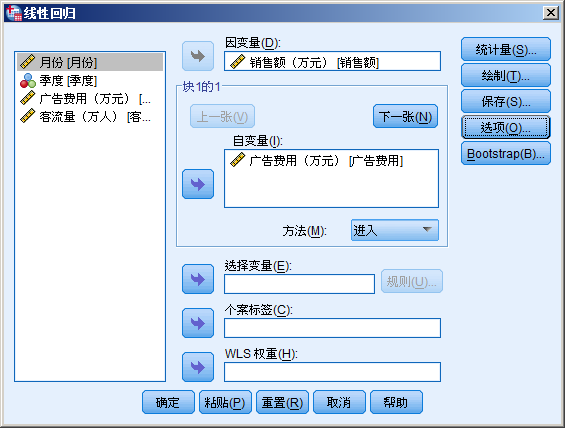
广告费用和销售额具有高度强相关，相关系数0.816，显著性差异小于0.05，具有统计学意义，广告费用越多，销售额越多。

4）建立线性回归模型

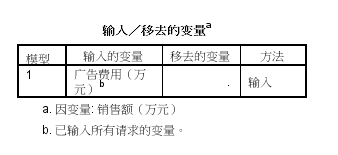








5）对模型进行检验



模型的方法

模型汇总表

看多大程度上能

对因变量做出解释

一元看R方

R方越大越好

多元看调整R方



1.R方的解释

R方在一元线性回归模型中，衡量【响应变量X和预测变量Y】的线性关系。

R方=cor（X,Y)^2

但是在多元线性回归模型中，因为涉及多个预测变量，所有R方就是**衡量响应变量和多个预测变量**之间的关系。

而相关系数，只是衡量一对变量之间的关系，所有就不能推广了。

R平方=cov（y，yi）^2

其中相关系数的两个变量变成，响应值和线性回归的预测值了。当然一元线性也同样适用了。

另一个公式  R-squared = 1-   (RSS/（n-p-1)) / (TSS/(n-1))

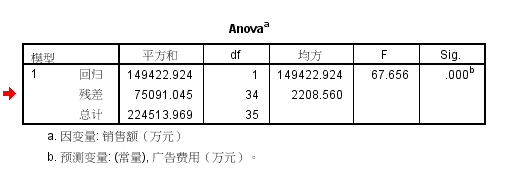
因为在模型中，增加多个变量，即使事实上无关的变量，也会小幅度条R平方的值，当时其是无意义，所有我们调整了下，降低R平方的值。

简单地说就是，用r square的时候，不断添加变量能让模型的效果提升，而这种提升是虚假的。

利用adjusted r square，能对添加的非显著变量给出惩罚，也就是说随意添加一个变量不一定能让模型拟合度上升

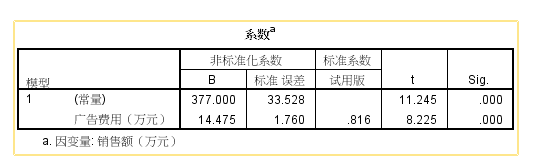
2.假设检验

一般而言，为了确定从样本(sample)统计结果推论至总体时所犯错的概率，我们会利用统计学家所开发的一些统计方法，进行统计检定。

https://www.cnblogs.com/Acceptyly/p/3930006.html

F值要查表，比较麻烦

Sig值（P值）越小越好  
（显著性）

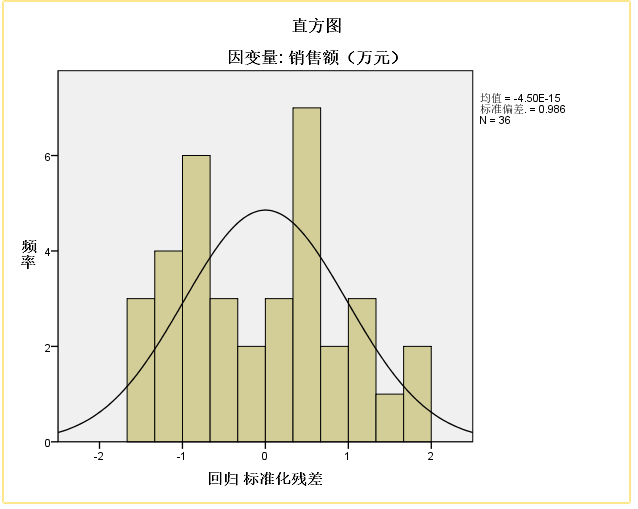


回归方程式

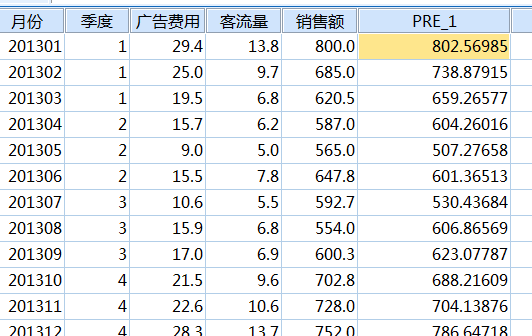
常量与回归系数

Y=377.000+14.475x



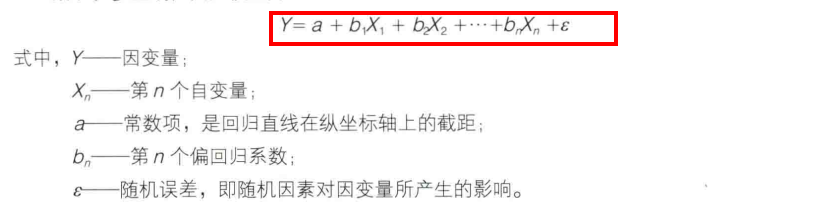


6）使用模型进行预测

### 【知识点8】SPSS多元线性回归分析

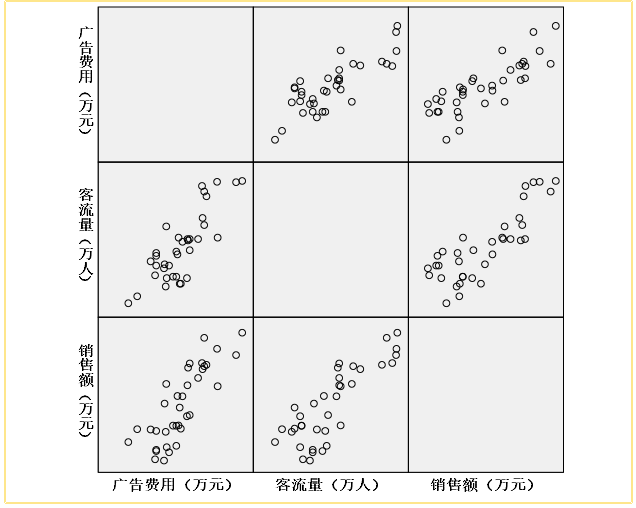
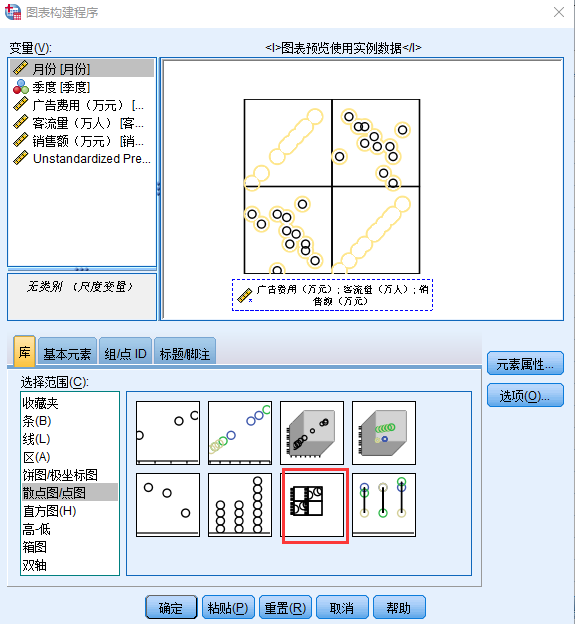
多元线性回归是指包含两个或两个以上自变量的线性回归模型，其模型为



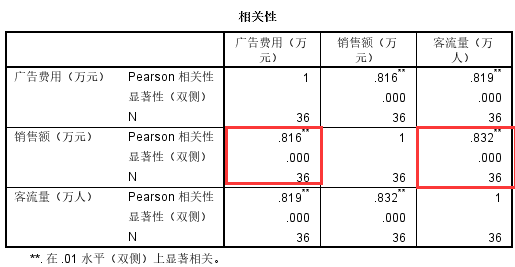
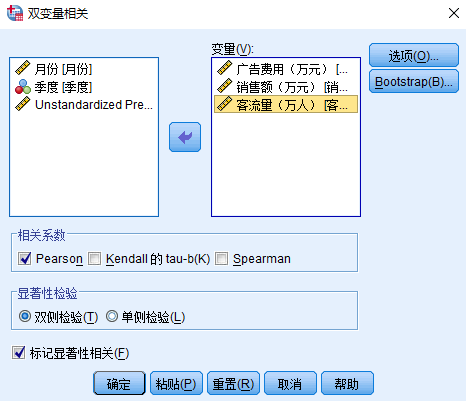
多元线性回归模型的建模过程和一元线性回归建模过程基本一致。

案例：对超市销售数据*.sav*建立线性回归模型，预测未来走势。

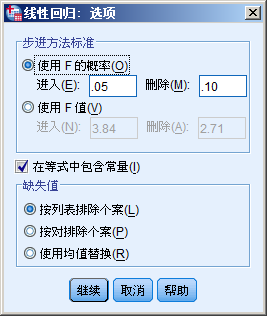
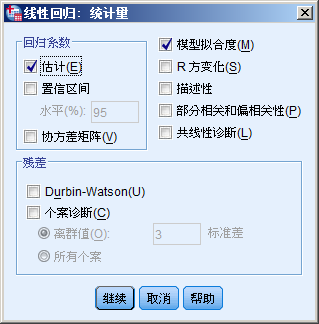
1、绘制散点图观察变量之间的关系



2、通过皮尔逊系数了解变量之间的相关关系

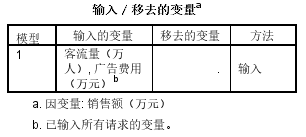


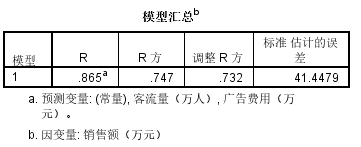
3、建立线性回归模型



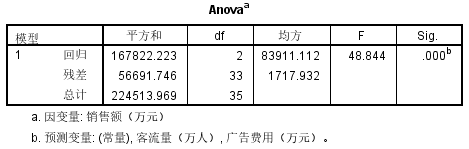


5、对模型进行检验

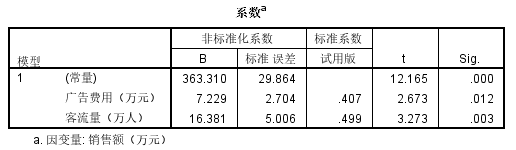
模型的方法



一元看R方，R方越大越好，多元看调整R方

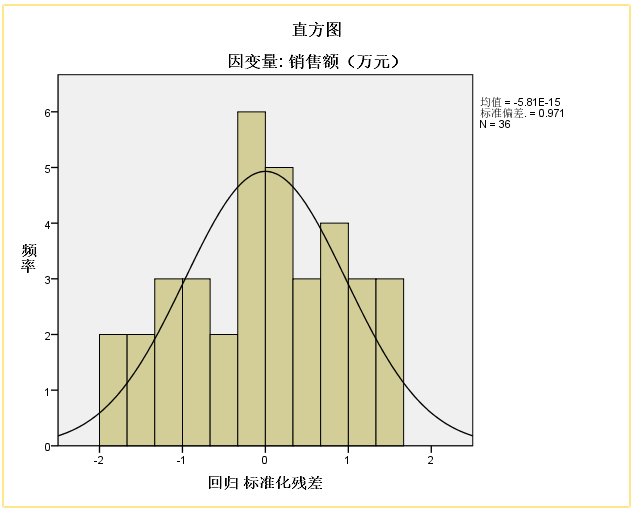
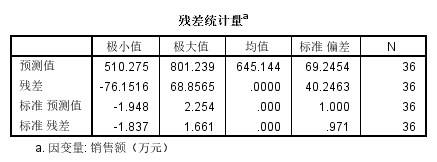


F值要查表，比较麻烦，Sig值（P值）越小越好（显著性）



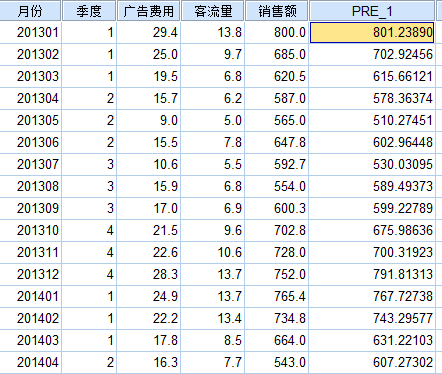
常量与回归系数

Y=363.310+7.229\*x1+16.381\*x2



客流量比广告费用提升销售额效果更好

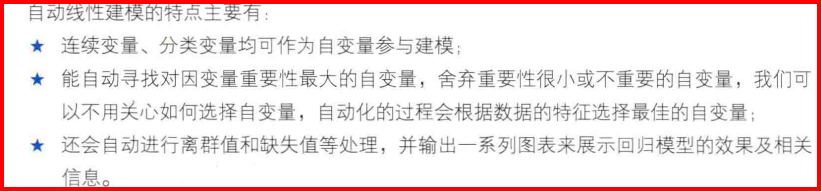
6、使用模型进行预测



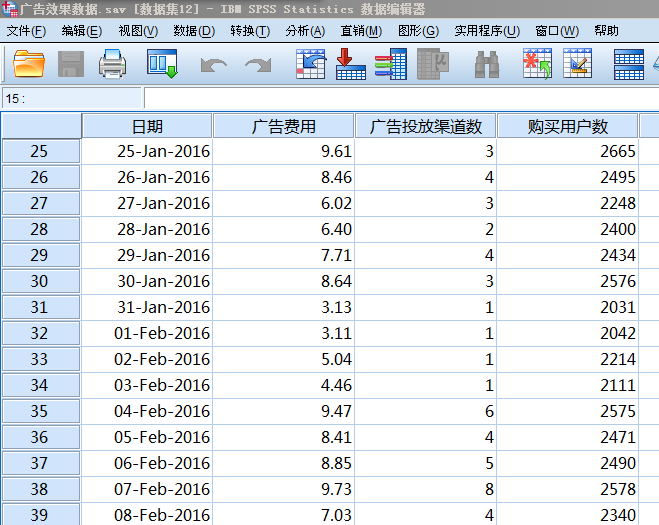
### 【知识点9】SPSS自动线性建模

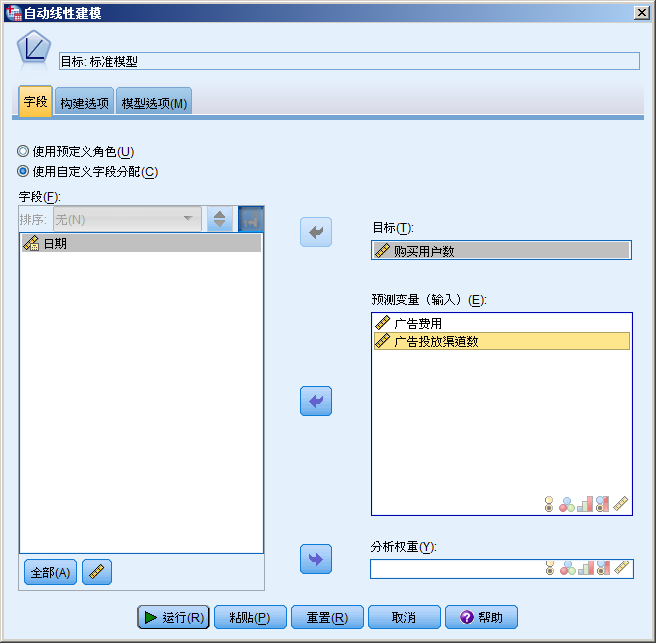
位置：分析-回归-自动线性建模

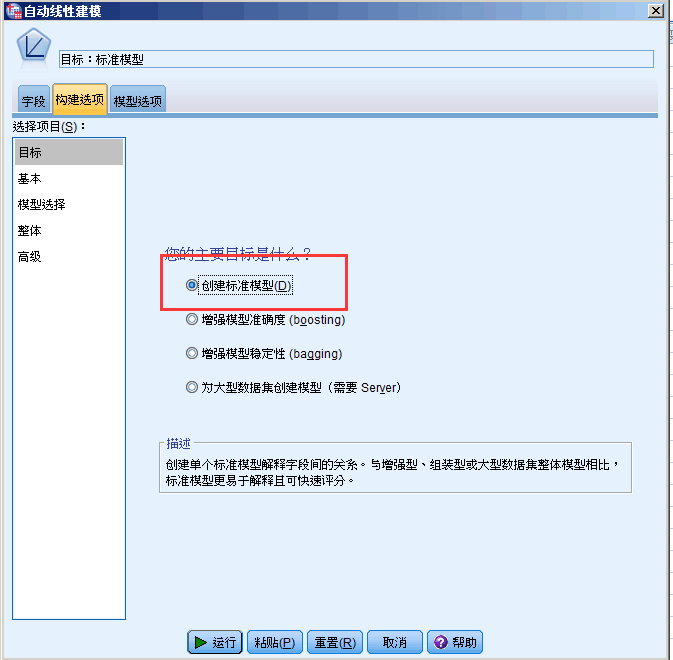
自动线性建模是在一般线性模型基础上加以改进，让用户输入最少的参数而自动建立线性模型

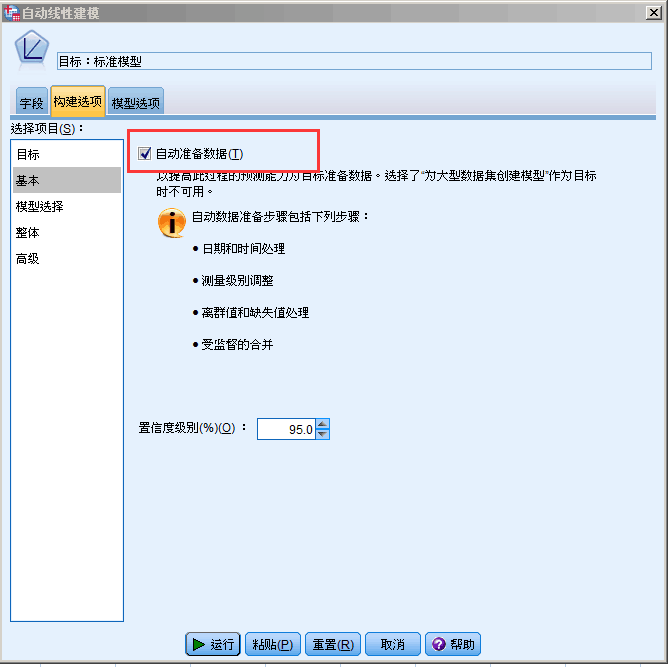


案例：对广告效果数据*.sav进行自动线性建模，*预测未来走势。



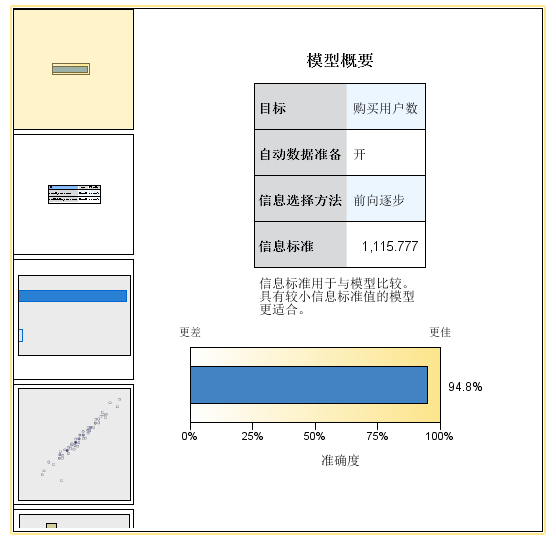




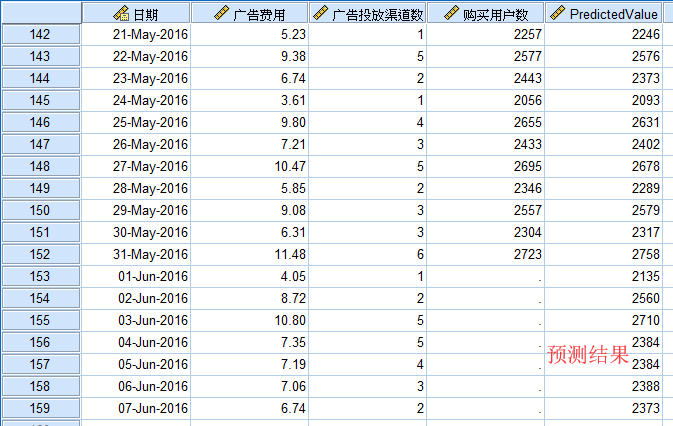




模型结果解读：



预测结果



## 自主学习作业讲解

无

## 课程总结

* SPSS数据预分析
* 频率分析
* 描述分析
* 探索分析
* 交叉列联表分析
* 相关分析
* SPSS回归分析
* SPSS多元线性回归分析
* SPSS自动线性建模

## 下次自主学习任务布置

#### 观看预习视频

数据建模分析师——SPSS数据分析

#### 课后作业

1. 参见直播平台
2. Qq群课后习题