现代统计分析方法与应用（第二版）

何晓群 中国人民大学出版社

，则拒绝原假设

第三章 定性数据的检验

3.1 多项分布与检验

一般要求样本量n比较大，使得每一类中的期望值不少于5.

例3.1 某超市为了研究顾客对三种品牌矿泉水的喜好比例，以便为下一次进货提供决策，随机观察了150名购物者，并记录下他们所买的品牌，统计出购买三种品牌的人数，如下表所示，这些数据是否说明顾客对这三种矿泉水的喜好确实存在差异？

|  |  |  |  |
| --- | --- | --- | --- |
| 品牌 | 甲 | 乙 | 丙 |
| 人数 | 61 | 53 | 36 |

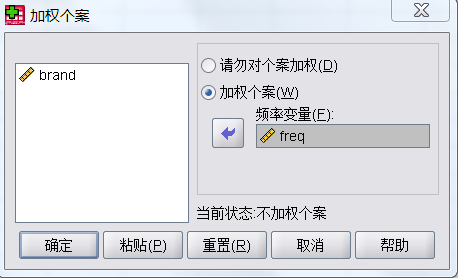
原假设顾客对三种矿泉水的喜好没有差异

备选假设顾客对三种矿泉水的喜好确实有差异

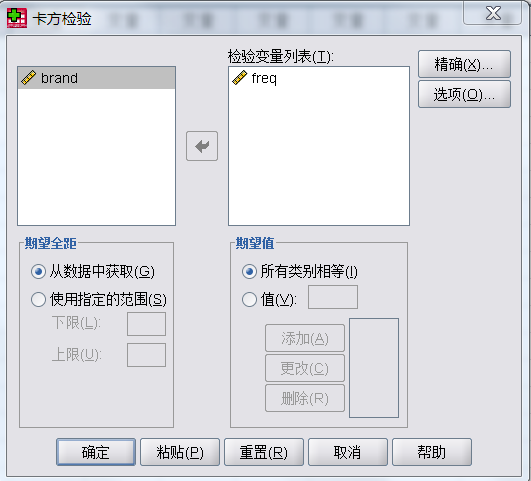
输入数据



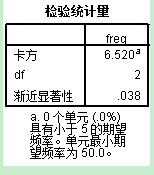
选择”数据”-”加权个案”



选择”分析”-”非参数检验”-”卡方”



确定，输出结果如下：



，拒绝原假设，

注：也可以是原假设概率分别取不同的值

3.2 列联表分析

一般要求样本量n比较大，使得每一类中的期望值不少于5.

例3.3 一个建筑装饰公司对一个城市三个地区的乔迁居民喜好地板的比例感兴趣，目的是决定对这些地区应采取何种营销策略。这个公司的调研部进行了一项调查，如下表，这些数据能否提供证据说明该城市不同地区的居民对地板喜好的比例确有不同？

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 地板材料 | 地区 | | | 合计 |
| 1 | 2 | 3 |
| 地板 | 69 | 126 | 16 | 211 |
| 其他 | 78 | 99 | 27 | 204 |
| 合计 | 147 | 225 | 43 | 415 |

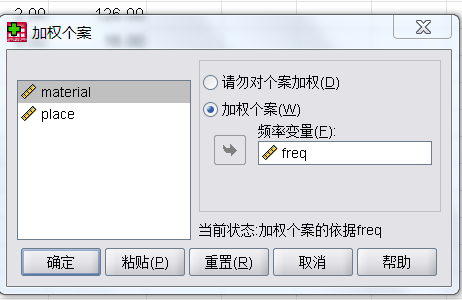
原假设居民对地板的喜好与地区无关（即两个变量互相独立）

备选假设居民对地板的喜好与地区有关

输入数据：



选择”数据”-”加权个案”

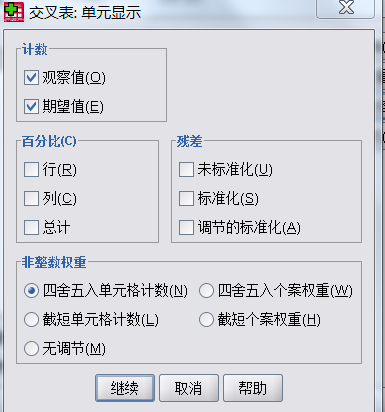


选择”分析”-”描述统计”-”交叉表”

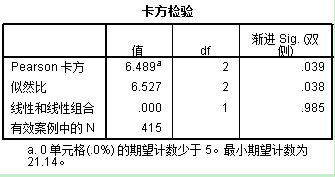


单击”统计量”按钮，选择”卡方”选项，”继续”

单击”单元格”按钮，选择”观察值”, “期望值”选项, “继续”

确定，输出结果如下：



，拒绝原假设，

3.3 一致性检验

例3.6 一个广告公司想知道n组年龄不同的人所喜欢的电视节目是否不同，他们在不同年龄的三组人中各抽选了一个随机样本，并请求样本中的每一个人回答在三类电视节目中他或她喜欢哪一类。调查结果如下表所示，括号内是预期频数。

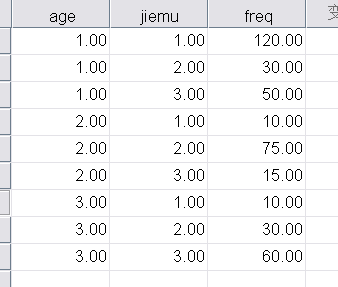
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 年龄小组总体 | 节目类型 | | | 合计 |
| A | B | C |
| 30岁以下 | 120 | 30 | 50 | 200 |
| 30~45岁 | 10 | 75 | 15 | 100 |
| 45岁以上 | 10 | 30 | 60 | 100 |
| 合计 | 140 | 135 | 125 | 400 |

一致性检验与独立性检验的差别在于抽样程序不同

就所喜欢的电视节目类型而言三个年龄组是一致的

三个年龄组不一致

SPSS操作同列联表分析的独立性检验





3.4 拟合度检验

总体分布检验，将样本的取值进行分类，可以进行适当合并，使每一类的期望频数等于或超过5。

例3.7 某宾馆为了研究每日注销的房间数所遵从的分布规律，管理人员观察在为期300天的时间内每日注销的房间数，如下表所示。这些数据是否同“每日注销的房间数遵从泊松分布”这一假设相容？

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 注销房间数 | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13及以上 |
| 观察到这一注销数的天数 | 3 | 9 | 21 | 38 | 46 | 54 | 49 | 34 | 20 | 16 | 6 | 3 | 1 | 0 |

原假设总体服从泊松分布

备选假设总体不服从泊松分布

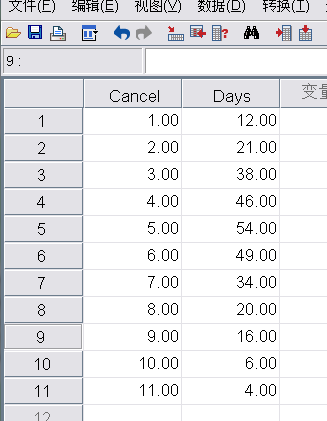
点估计参数

合并数据，后三个数据合并为 11个及以上有4天，前二个数据合并为不超过1个有12天

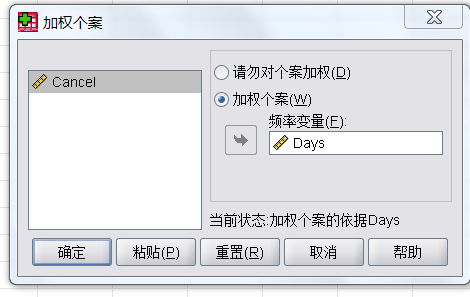
利用EXCEL计算各类的理论频数如下表所示：

|  |  |  |  |
| --- | --- | --- | --- |
| 天数 | 频数 | 理论频率（参数5.2） | 理论频数 |
| <=1 | 12 | 0.034202699 | 10.26081 |
| 2 | 21 | 0.074583951 | 22.37519 |
| 3 | 38 | 0.129278848 | 38.78365 |
| 4 | 46 | 0.168062503 | 50.41875 |
| 5 | 54 | 0.174785003 | 52.4355 |
| 6 | 49 | 0.151480336 | 45.4441 |
| 7 | 34 | 0.11252825 | 33.75847 |
| 8 | 20 | 0.073143362 | 21.94301 |
| 9 | 16 | 0.042260609 | 12.67818 |
| 10 | 6 | 0.021975517 | 6.592655 |
| >=11 | 4 | 0.017698922 | 5.309677 |
| 合计 | 300 | 1 | 300 |

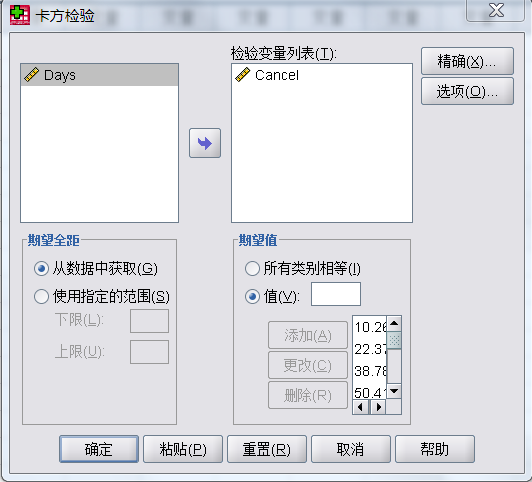
SPSS中输入数据



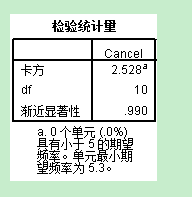
选择”数据”-”加权个案”



选择”分析”-”非参数检验” -”卡方”，把Excel中算的的期望频数输入到期望值中



确定，输出结果如下



，没有充分理由拒绝原假设，可以认为服从泊松分布

第四章 一元线性回归

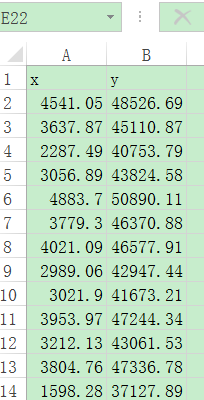
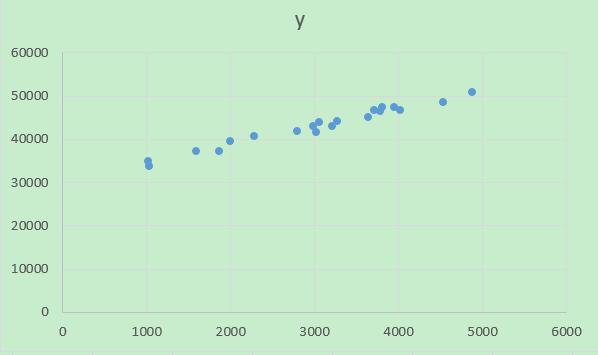
研究某一现象与影响它的某一最主要因素的影响，

例4.1 假定我们需要研究化肥施用量与粮食产量的关系，以便确定合理的化肥施用量。如下表列出了20组数据。

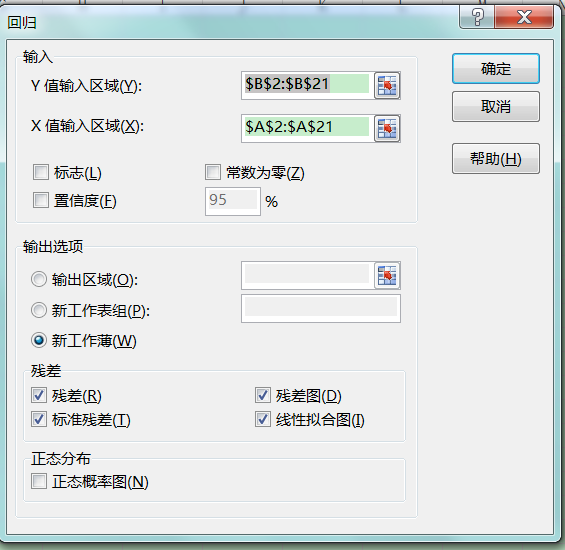
|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 化肥使用量x（万吨） | 4541.05 | 3637.87 | 2287.49 | 3056.89 | 4883.7 | 3779.3 | 4021.09 |
| 粮食产量y（万吨） | 48526.69 | 45110.87 | 40753.8 | 43.824.58 | 50890.11 | 46370.88 | 46577.91 |
| 化肥使用量x（万吨） | 2989.06 | 3021.9 | 3953.97 | 3212.13 | 3804.76 | 1598.28 | 1998.56 |
| 粮食产量y（万吨） | 42947.44 | 41673.21 | 47244.3 | 43061.53 | 47336.78 | 37127.89 | 39515.07 |
| 化肥使用量x（万吨） | 3710.56 | 3269.03 | 1017.12 | 1864.23 | 2797.24 | 1034.09 |  |
| 粮食产量y（万吨） | 46598.04 | 44020.92 | 34866.9 | 37184.14 | 41864.77 | 33717.78 |  |

输入EXCEL，做散点图，观察基本线性，可以用尝试一元线性回归，如果散点图非线性，可能要尝试非线性回归

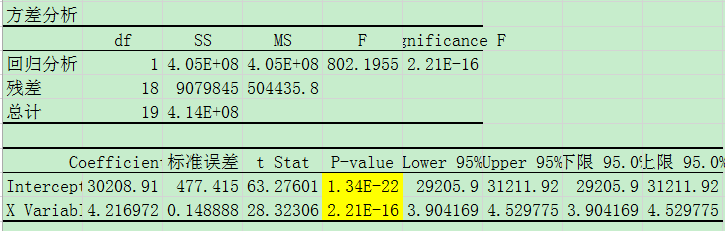
一元线性回归模型： 其中

加载Excel数据分析功能，工具-加载宏-分析工具，在工具下拉菜单中就增加了数据分析项（不同版本EXCEL可能有不同），点选数据分析项，选择回归，

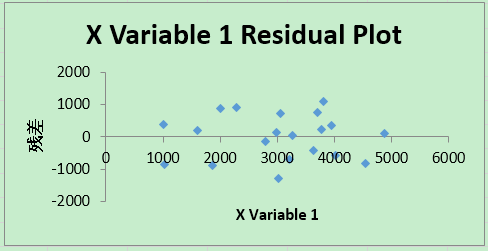


输出结果如下



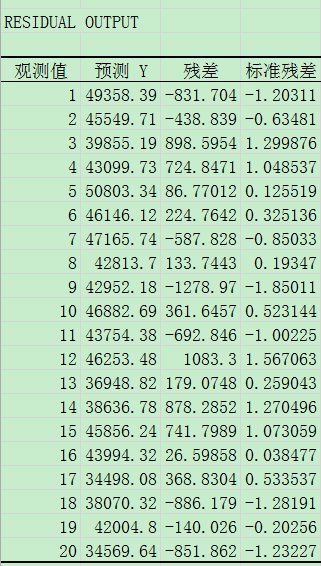
上图黄色部分即为常数项和一次项系数的显著性检验p值，明显，拒绝原假设，故和都显著不为0，回归方程

注意，我们主要关心的是的显著不为0，代表y随着x的变化而变化，但在利用回归方程分析和预测之前，应该用残差图帮助我们诊断回归效果与样本数据的质量，检查模型是否满足基本假定



如上残差图在0上下无规律波动，说明回归模型满足基本假设，

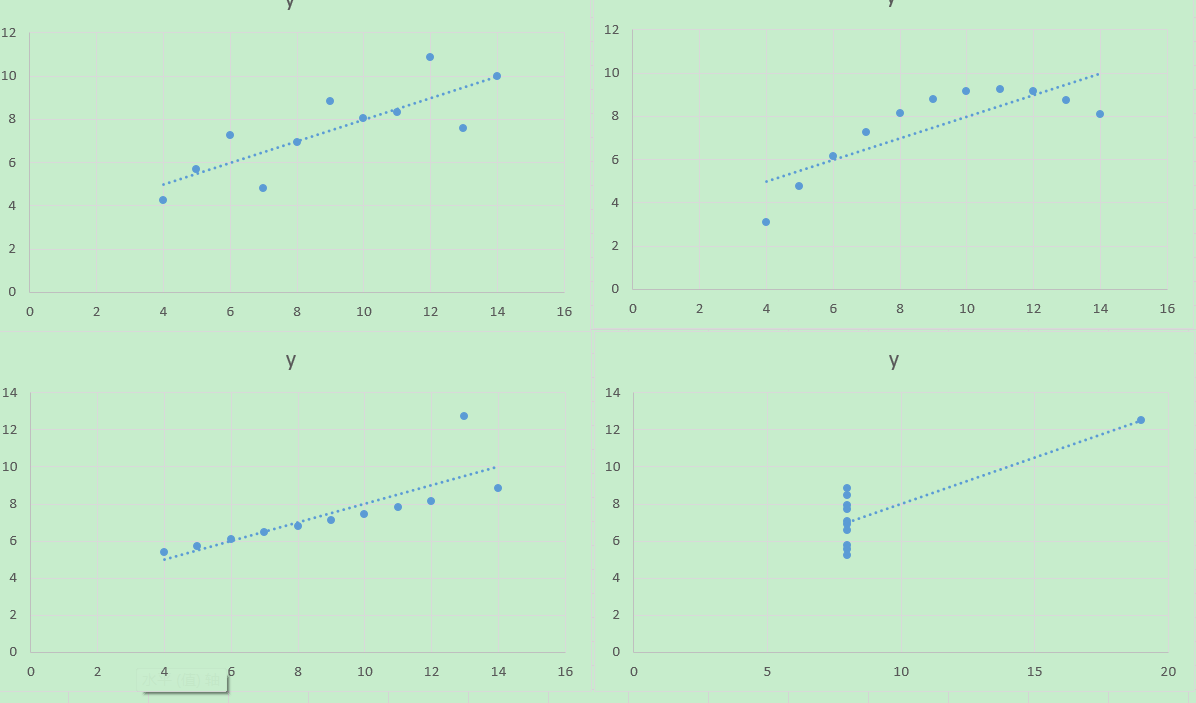
在残差分析，还应该注意删去异常值，标准残差的绝对值>3的对应观测值为异常值，应删去



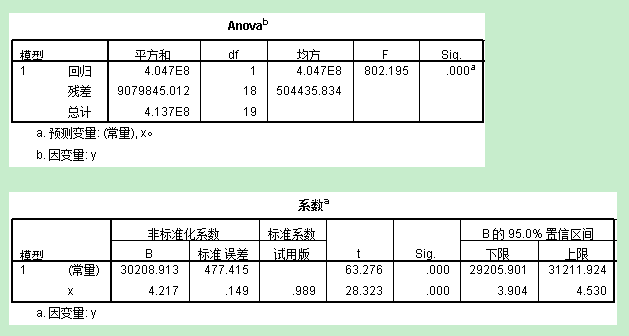
做残差分析的理由如下图，四组数据线性回归结果相同，且都通过显著性检验，但从拟合图形上很明显看出，图一的线性回归是合适的；图二应该用非线性回归可以提高决定系数，判定系数取值[0,1]之间，越接近1说明回归效果越好；图三明显存在异常值；图四缺少中间数据，应补充数据。

这个例子告诉我们，但拒绝原假设时，我们说y和x之间存在线性相关性，但是并不能肯定的说线性相关性就是最好的描述，应综合决定系数、散点图、残差分析等综合分析。

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 第一组 | | 第二组 | | 第三组 | | 第四组 | |
| x | y | x | y | x | y | x | y |
| 4 | 4.26 | 4 | 3.1 | 4 | 5.39 | 8 | 6.58 |
| 5 | 5.68 | 5 | 4.74 | 5 | 5.73 | 8 | 5.76 |
| 6 | 7.24 | 6 | 6.13 | 6 | 6.08 | 8 | 7.71 |
| 7 | 4.82 | 7 | 7.26 | 7 | 6.44 | 8 | 8.84 |
| 8 | 6.95 | 8 | 8.14 | 8 | 6.77 | 8 | 8.47 |
| 9 | 8.81 | 9 | 8.77 | 9 | 7.11 | 8 | 7.04 |
| 10 | 8.04 | 10 | 9.14 | 10 | 7.46 | 8 | 5.25 |
| 11 | 8.33 | 11 | 9.26 | 11 | 7.81 | 8 | 5.56 |
| 12 | 10.84 | 12 | 9.13 | 12 | 8.15 | 8 | 7.91 |
| 13 | 7.58 | 13 | 8.74 | 13 | 12.74 | 8 | 6.89 |
| 14 | 9.96 | 14 | 8.1 | 14 | 8.84 | 19 | 12.5 |



利用SPSS回归的方法也非常简单，输入数据后，选择”分析”-”回归”-”线性”，进入线性回归窗口，输出如下



sig即为显著性检验的p值，明显，拒绝原假设，故和都显著不为0，

在SPSS软件的线性回归对话框中，点选Save，可以选择保存所需中间变量，包括标准化残差和学生化残差

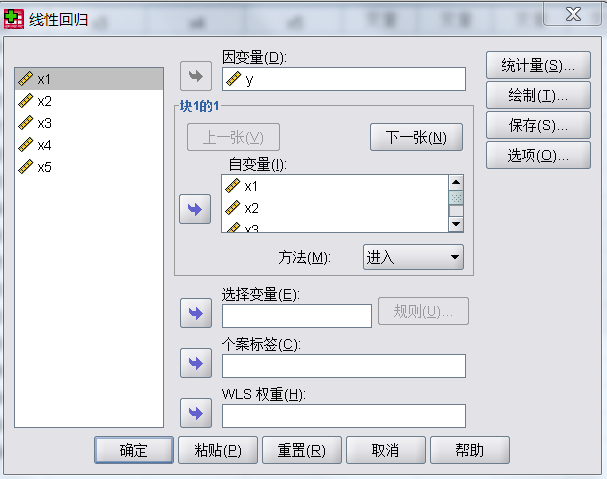
第五章 多元线性回归

多元线性回归模型：

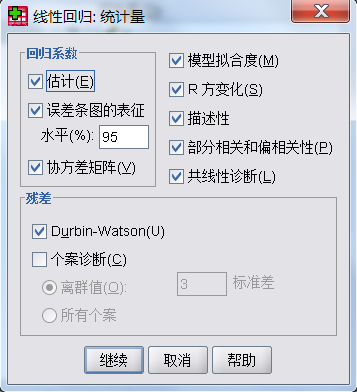
例5.3 中国民航客运量的回归模型，为了研究我国民航客运量的变化趋势及其成因，我们以民航客运量作为因变量y（万人），以国民收入x1（亿元）、消费额x2（亿元）、铁路客运量x3（万人）、民航航线里程（万公里）、来华旅游入境人数（万人）。根据《1994年统计摘要》获得数据如下表所示：

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| year | y | x1 | x2 | x3 | x4 | x5 |
| 1978 | 231 | 3010 | 1888 | 81491 | 14.89 | 180.92 |
| 1979 | 298 | 3350 | 2195 | 86389 | 16 | 420.39 |
| 1980 | 343 | 3688 | 2531 | 92204 | 19.53 | 570.25 |
| 1981 | 401 | 3941 | 2799 | 95300 | 21.82 | 776.71 |
| 1982 | 445 | 4258 | 3054 | 99922 | 23.27 | 792.43 |
| 1983 | 391 | 4736 | 3358 | 106044 | 22.91 | 947.7 |
| 1984 | 554 | 5652 | 3905 | 110353 | 26.02 | 1285.22 |
| 1985 | 744 | 7020 | 4879 | 112110 | 27.72 | 1783.3 |
| 1986 | 997 | 7859 | 5552 | 108579 | 32.43 | 2281.95 |
| 1987 | 1310 | 9313 | 6386 | 112429 | 38.91 | 2690.23 |
| 1988 | 1442 | 11738 | 8038 | 122645 | 37.38 | 3169.48 |
| 1989 | 1283 | 13176 | 9005 | 113807 | 47.19 | 2450.14 |
| 1990 | 1660 | 14384 | 9663 | 95712 | 50.68 | 2746.2 |
| 1991 | 2178 | 16557 | 10969 | 95080 | 55.91 | 3335.65 |
| 1992 | 2886 | 20223 | 12985 | 99693 | 83.66 | 3311.5 |
| 1993 | 3383 | 24882 | 15949 | 105458 | 96.08 | 4152.7 |

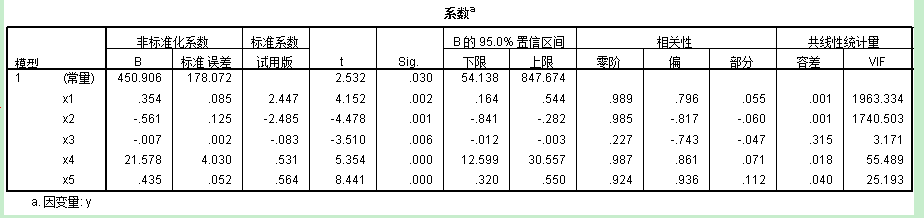
将数据输入SPSS中，选择”分析”-”回归”-”线性”



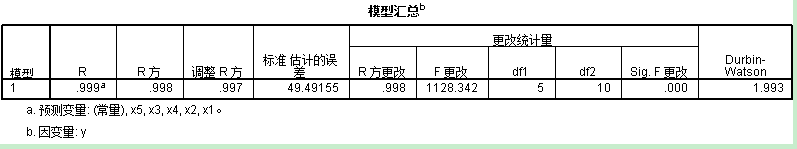
选择“统计量”按钮，在弹出选项卡中选择选项



输出结果如下

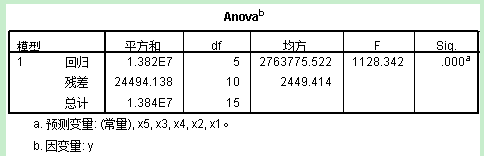


得到原始回归模型如下：

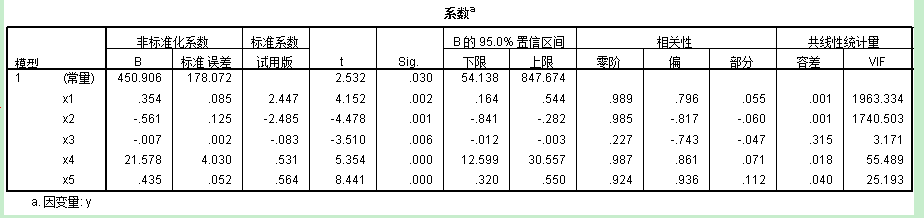


由上表可以看出，决定系数0.998，回归方程高度显著

杜宾沃森检验值DW=1.993表明无自相关性，（DW在2附近最好）



方差分析表可见，F=1128.342，p值=0.000很小，表明回归方差高度显著



回归系数的显著性检验，5个自变量的p值都很小，说明自变量都对y均有显著影响。

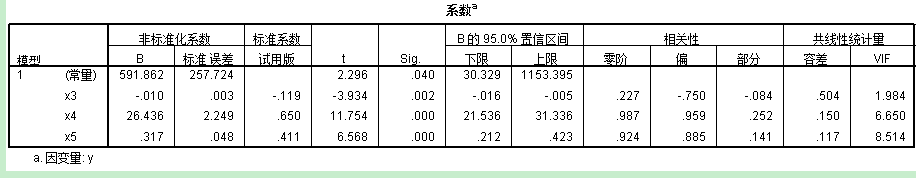
多重共线性检验，上面系数表中的方差膨胀因子VIF可以判断变量间是否存在多重共线性，一般认为VIF>10时该变量在模型中产生共线性，共线性的产生会对回归预测产生严重的影响。可以逐步剔除一些VIF大的变量，重新进行回归，直到所有VIF值都小于10为止。

本例中，依次剔除后，再回归结果符合要求，

回归方程为：

标准化后回归方程为：

由标准化后回归系数可知，对民航客运量影响最大的因素是民航航运里程，每增加1%，民航客运量会增加0.65%。



第10章 聚类分析

聚类分析的目的就是把相似的研究对象归成类