**一 数据情况描述**

鉴于1169条原始数据中含有很多缺少平均分数线或者其他重要信息的数据，首先对其进行初步的数据清洗，最终符合条件的数据共有411条。

首先我们对其中的变量进行初步的描述性统计说明：



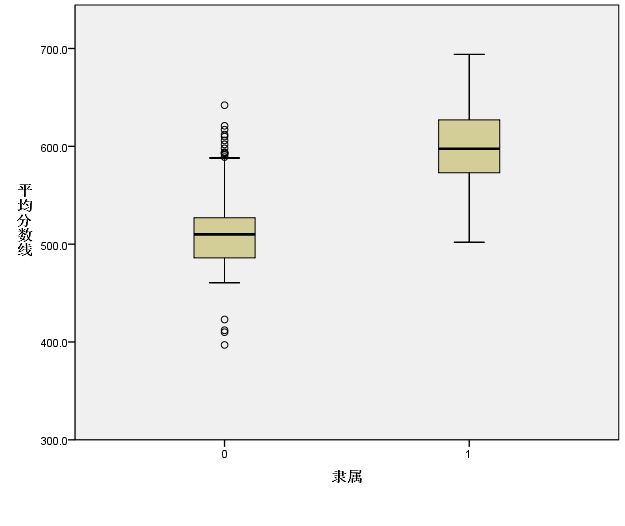
表 1：高校录取平均分数线变量数据说明

**二 文本型数据数值化**

鉴于目的是构造一个预测型回归模型，所以有必要将文本型解释变量转换为数值型，对于***是否为211，是否为985，是否自主招生*** 这三个二分类变量采取将“是”标值为1，“否”标值为0的方法即可解决这个问题。

对于***所在地，隶属*** 以及***院校类型*** 这三个文本型数据的数值化需要做进一步处理：

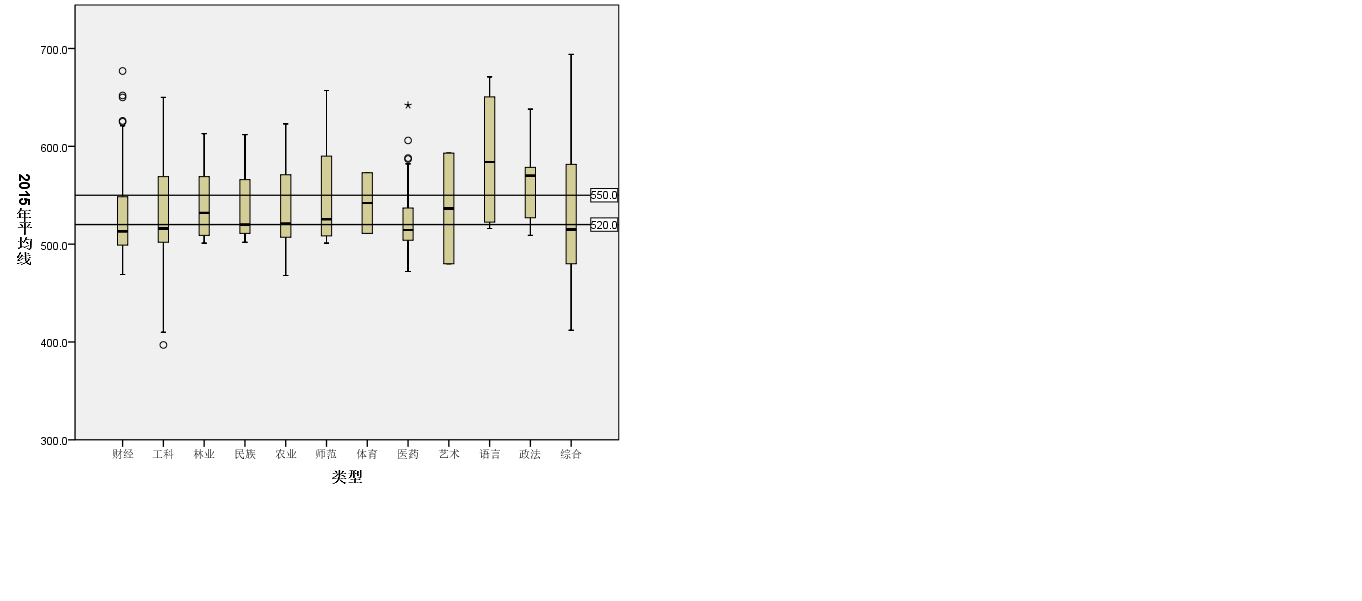
1 隶属

 根据教育部或公安部等全国性部属高校的受热捧程度普遍高于普通市教育厅属高校的现实情况，将***隶属*** 为部属的标值为1，非部署的标值为0。

根据已有的设定，最终得出的箱线图也可以看到数据的展现内容确实和我们最初的感觉相符：部属高校的平均分数线普遍高于非部属高校。

图 1：2015年高校平均分与隶属关系箱线图

2 院校类型



通过观察发现语言与政法类大学的平均成绩明显高于普通高校，体育艺术类院校的成绩也普遍偏高。

检查原始数据发现，这几类院校平均分高于其他是由于经数据筛选后几类院校的数据量极少--甚至只有1个，不具参考意义；又由于其他类院校所占比重超过95%且平均分并未显示较大差异，所以最 终决定将院校类型这个变量予以剔除。

图 2：2015年高校平均分数线与学校类型箱线图

3 所在地

如表2所示，根据现有的对中国科教之城市等级划分划分标准，可以对各个城市标记等级，等级跨度如图表所示，对于级别跨度大的城市之间跃度定为2，跨度小的跃度定为1，这样的赋值比较符合现实世界的情况，最终依据得到的1~10的城市等级量化表来对高校***所在地*** 文本型变量进行数值化处理。



表 3：城市等级量化表[1]

**三 逐步回归分析（预测型回归）[2]**

预测型回归的目的是通过变量选择程序来建立一个最佳的预测方程，以用于实际的预测分析。为了选择最佳独立变量组合，此处使用逐步法，以逐步回归方法决定最佳模型。



逐步回归法中独立变量的进入或删除清单，以及选择标准。进入以F概率.05，删除以F概率.10为标准

总计六个变量分八个步骤被选入回归方程式----是否为211，所在地，隶属，重点学科数，硕士点，是否为985.

根据模型汇总及变异数分析，模型显著性整体检验（其中模型8的R2=.649，F检验值124.447）[详见附件]均达显著，表示回归效果具有统计意义。

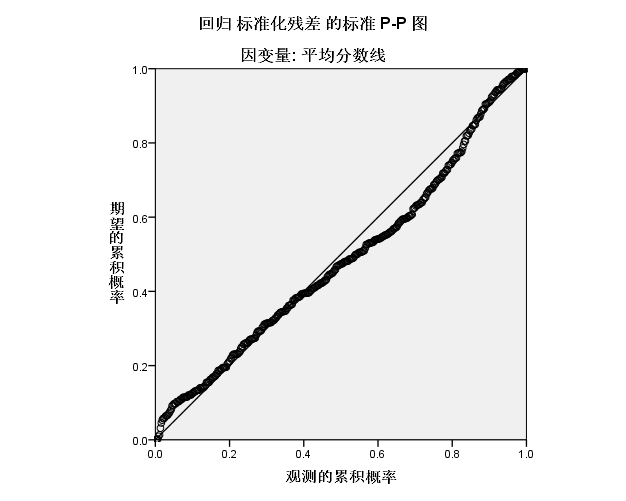
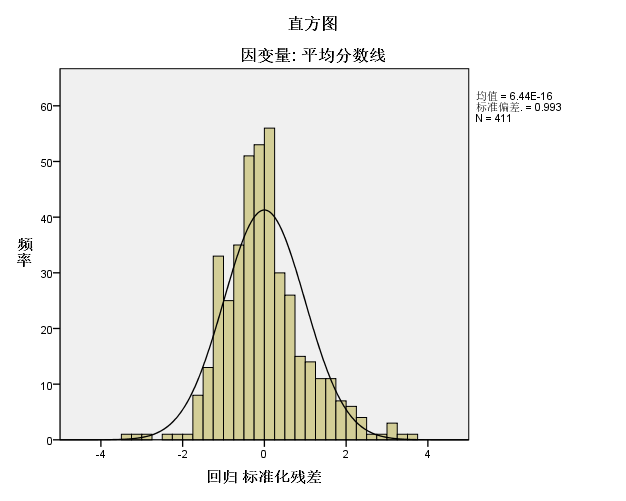
逐步参数估计

进入模型8的各个变量的系数求得如下，T检验均达显著，无共线性问题

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 非标准化系数 | | 标准系数 | t | Sig. | 相关性 | | | 共线性统计量 | |
| B | 标准误差 | 试用版 | 零阶 | 偏 | 部分 | 容差 | VIF |

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 8 | (常量) | 486.469 | 3.606 |  | 134.890 | .000 |  |  |  |  |  |
| 是否为211 | 42.624 | 6.588 | .346 | 6.470 | .000 | .737 | .306 | .191 | .304 | 3.292 |
| 所在地 | 3.964 | .594 | .213 | 6.677 | .000 | .441 | .315 | .197 | .852 | 1.174 |
| 隶属 | 25.183 | 6.115 | .204 | 4.119 | .000 | .684 | .201 | .121 | .353 | 2.836 |
| 重点学科(个)  硕士点(个) | .410  .085 | .202  .032 | .071  .159 | 2.034  3.926 | .043  .000 | .414  .537 | .101  .219 | .060  .089 | .720  .587 | 1.389  1.502 |
| 是否为985 | 15.866 | 7.193 | .087 | 2.206 | .028 | .547 | .109 | .065 | .554 | 1.807 |

标准化残差分布如下图，基本符合正态分布，验证了模型的科学性，该模型具有统计意义。



**四 结果说明**

逐步回归模型以逐步分析法选择最佳独立变量组合，计算机自动选取相关最高的独立变量首先进入模型，可以暂时回避共线性问题。最终的模型八共有*是否为211、所在地、隶属、重点学科数、硕士点、是否为985* 六个解释变量，总计可解释因变量64.9%的变异量，调整后为64.4%。根据F检验结果，此一解释力具有统计意义（F=124.447，p=.000）。最后得到的方程如下：

利用这个方程式，可以进行实际的高校分数线预测，估计标准误为31.7946。

**【参考文献】**

[1] 王言荣;刘洁. 中国城市科教职能等级划分及空间分布研究

[2] 邱皓政.《量化研究与统计分析—SPSS中文视窗版数据分析范例解析》

【附件】

