1. 数据预处理
2. 查看缺失值情况
3. 对缺失值进行处理

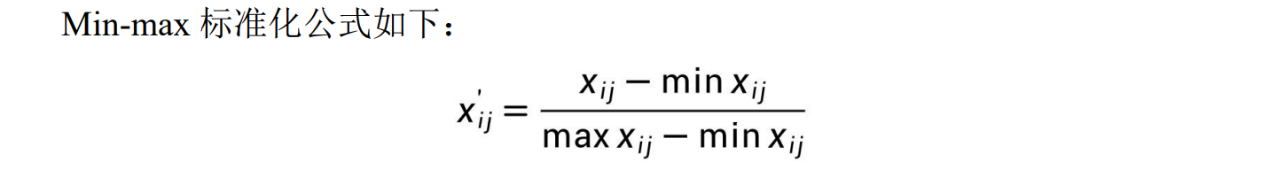
本文对数值型数据使用均值进行填补,对分类型数据使用众数补全.

1. 进行异常值处理

本文使用Z-score方法来识别和处理异常值，(原理)

1. 数据标准化

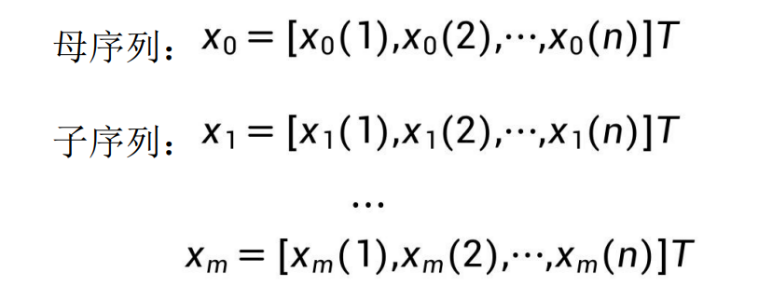
不同种类的的数据单位不统一将导致比较失去意义，由此我们需要借助数据标准化的方式是两者居于同一比较地位，相对于零均值标准化，我们倾向于采用 min-max 标准化对数据进行归一化处理，将表格中的数据经过变换转化成没有量纲的表达式,缩放到0和1之间；目的是使各个特征维度对目标函数的影响权重是一致的；改变了原始数据的一个分布,

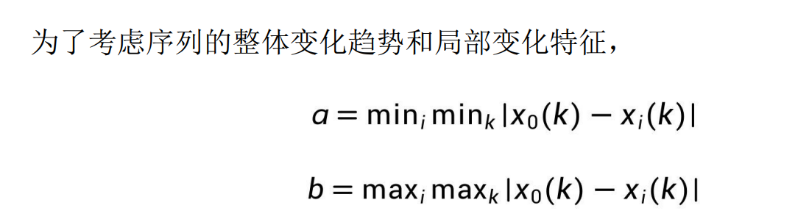


1. 灰色关联度分析

我们将AQI的值作为重要的评判指标,也就是自变量。被研究对象即因变量作为参考序列以反映空气质量等级的评判标准，对因变量有影响的各指标即 21 个自变量作为比较序列，分别用 X0 和 Xi（i=1, 2, …, 21）

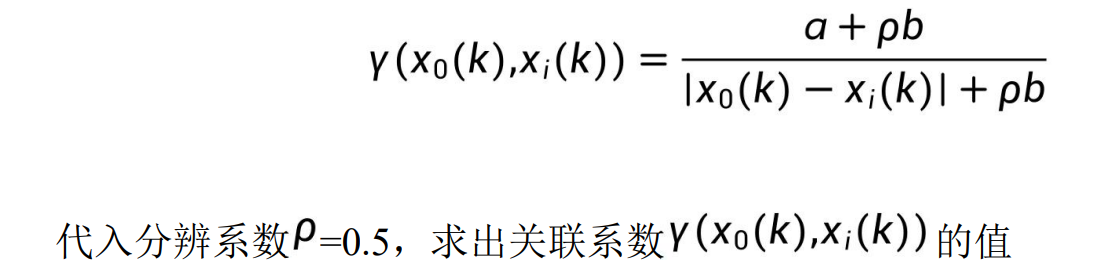
由于分析序列过分冗长，在此不做具体展示，对比子母序列并结合下列计算公式，



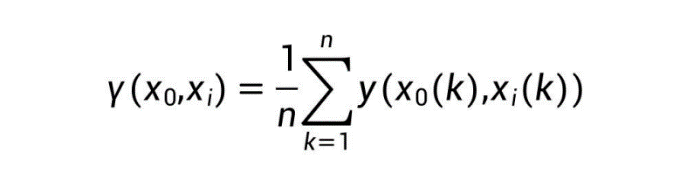


通过上述公式，我们可以计算出两级最小差值（a）和两级最大差值（b）。通过综合考虑最大值和最小值，我们能够更全面地描述序列的特征和变化趋势，并进而准确计算关联度。使用这种方法，可以在一定程度上平衡整体趋势和局部特征之间的关系，从而提高灰色关联模型的准确性和可靠性。

因此，我们将其代入最终的关联系数计算公式：



灰色关联度计算

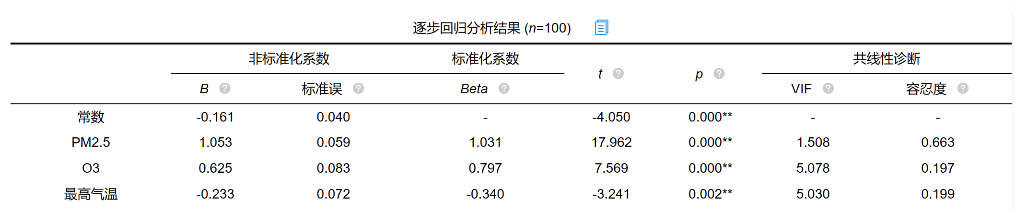


通过除以样本量 n，可以对关联系数进行归一化。由于元素组成和粒径所反应的每个数据点上的关联系数取值范围可能存在差异，并且样本量的大小也可能会影响计算结果的尺度。为了使得不同样本量和不同数据范围的序列能够进行比较和对比，我们需要对关联系数进行归一化处理，将其限制在[0,1]的范围内。

1. 逐步回归分析

逐步回归分析方法的是自动从21种可供选择的变量中选取最重要的变量，首先将自变量逐个引入，引入的条件是其偏回归平方和经检验后是显著的。同时每引入一个新的自变量后，要对旧的自变量逐个检验，剔除偏回归平方和不显著的自变量。这样一直边引入边剔除，直到既无新变量引入也无旧变量删除为止。它的实质是建立“最优”的多元回归方程。本文使用的是向前法建立回归分析的预测或者解释模型；

1. 首先分析模型拟合情况R方,以及可对VIF值或者容忍度值，容忍度=1/VIF值进行分析判断，VIF>5一般说明特征间有共线性



2.分析X的显著性,如果显著,则说明X对Y有影响关系,接着具体分析影响关系方向；

将PM2.5, PM10, CO, NO2, SO2, O3, 平均气温, 最高气温, 最低气温, 平均相对湿度, 日照时效, 降水量, 空气最差, 空气最好, 一日最大降水量, 平均风速, 平均空气质量指数, 平均高温, 平均低温, 极端高温, 极端低温作为自变量，而将AQI作为因变量进行逐步回归分析，经过模型自动识别，最终余下NO2, SO2, PM2.5, PM10, CO, 降水量, 平均风速, 最高气温作为自一共8种特征,最后得到模型的R方值为0.790，意味着这几种特征可以解释AQI的79.0%变化原因。并且模型通过F检验(F=120.716，p=0.000<0.05)，说明模型有效。

1. 有序逻辑回归

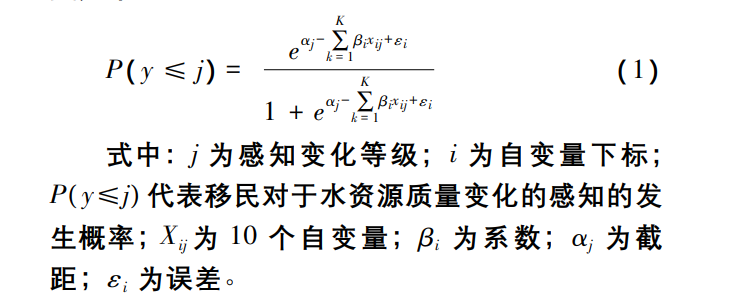
而有序回归适用于质量程度的因变量分析，模型是将逻辑分布作为随机误差项的概率分布的一种多元离散选择模型，适用于对按照效用最大化原则所进行的选择行为的分析。有序回归中 logit 的处理方法可以将一系列含固有非线性关系的因变量和自变量转化成一个线性的因变量和自变量的关系

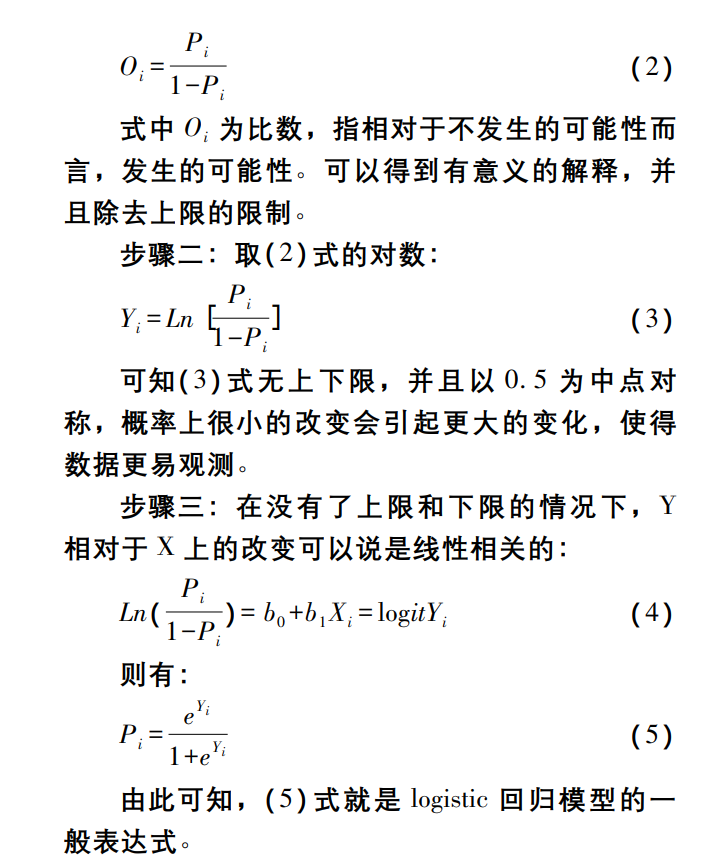
通过前面的相关性分析,我们决定将NO2, SO2, PM2.5, PM10, CO, 降水量, 平均风速, 最高气温这8种特征作为自变量，而将质量等级作为因变量进行有序logistic回归分析，并且使用Logit连接函数来预测当地某一时刻的天气状况属于哪一种质量等级。

1. 连接函数选择

因为选取的8种变量的分布较为均匀,所以我们选用Logit函数作为连接函数

原理





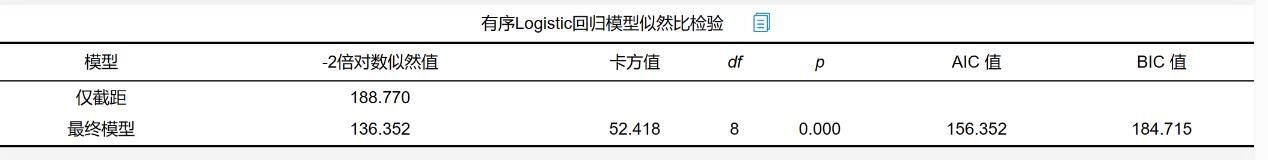
1. 平行性检验

平行性是有序Logit回归的前提条件，如果不满足平行性就无法使用该模型,因此进行平行性检验.

从结果可知p值大于0.05,说明模型接受原假设，即模型满足平行性检验,接受原假设。

1. 模型似然比检验

首先对模型整体有效性进行分析(模型似然比检验)

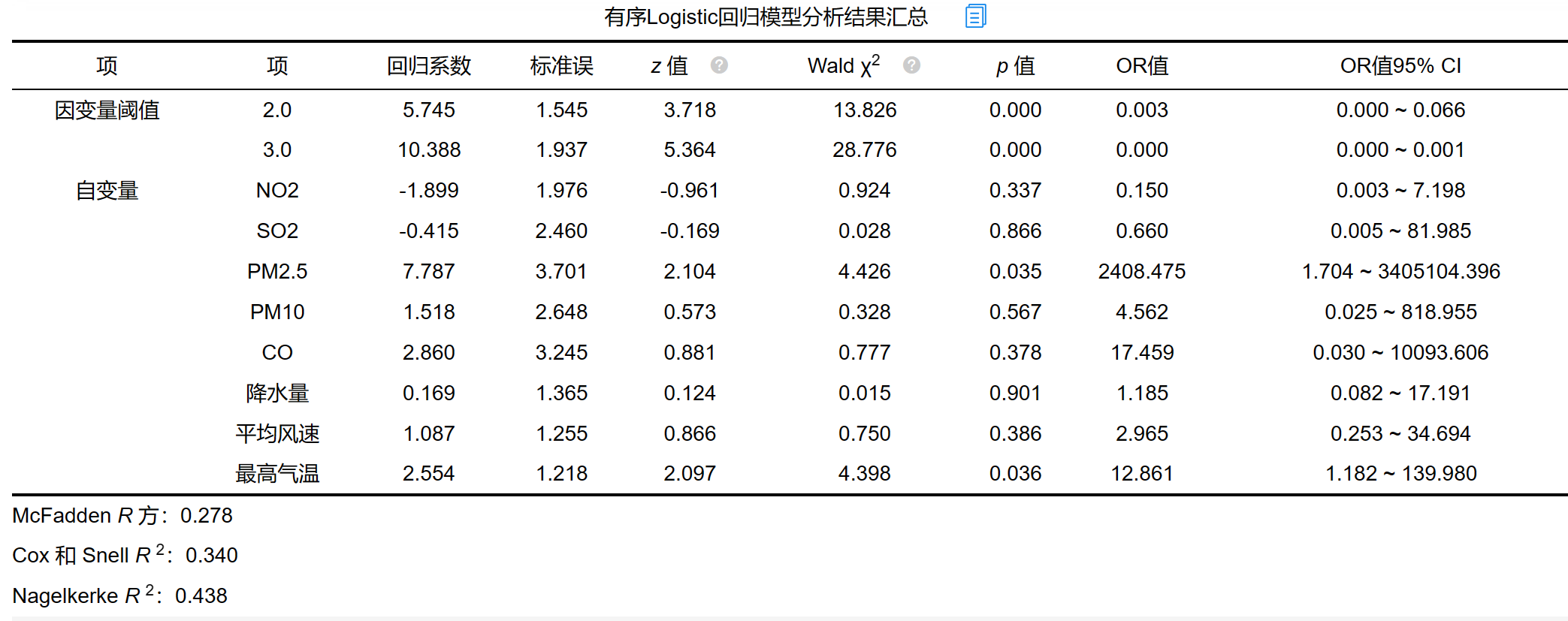


从上表可知：chi=52.418,p=0.000<0.05此处拒绝原假设,且说明本次构建模型时，放入的自变量具有有效性，本次模型构建有意义。

4} 模型建立和参数假设

参数估计原理:最大数似然估计

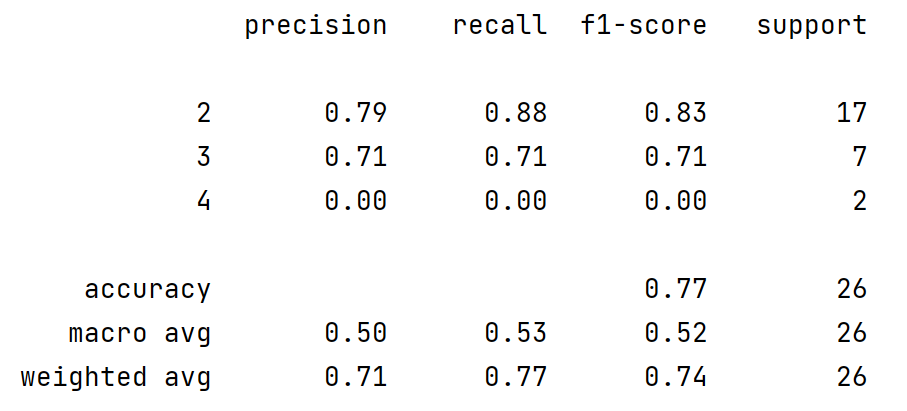
参数估计方法:梯度下降算法



得到有序逻辑回归模型

Logit(odds)=5.745-1.899\*NO2-0.415\*SO2+7.787\*PM2.5+1.518\*PM10+2.860\*CO+0.169\*降水量+1.087\*平均风速+2.554\*最高气温

1. 机器学习-有序逻辑回归算法



最后使用有序逻辑回归算法进行预测, 通过模型预测准确率去判断模型拟合质量，从上表可知：研究模型的整体预测准确率为77%，模型拟合情况可以接受.