**线性回归分析案例：**

# 城市铅暴露与婴儿死亡率的相关性分析

# 一、案例背景

铅（一种金属化学元素，元素符号Pb）是一种众所周知的有害物质，其不良影响主要包括增加婴儿死亡率、导致人类智商下降、发病率增加以及更加暴力等。多年来，不少学者为了研究铅暴露对于婴儿死亡率的影响，运用了各种各样的方法，但是很难得到一个较好的结果。其主要原因是，当前婴儿接触铅的主要途径是含铅汽油中的铅涂料以及空气中的铅，而这类数据通常难以获得并被用于分析。

而在十九世纪末和二十世纪初的美国，情况则完全不同，铅的暴露量不仅巨大，而且居民接触铅的渠道也不一样，主要通过水资源接触铅（即水源性铅暴露）。这是因为在上个世纪的美国各城市中，铅管被广泛用于连接住宅和建筑物以及街道干管。《美国自来水厂手册》报告提供的供水管道类型的数据显示，1897年，美国有67%的城市均使用了含铅管道。由于自来水可以溶解管道内壁的铅，使水中铅的含量飙升，从而导致居民大量地接触铅。1900年，美国仍有42%的城市全部使用铅管，25%的城市混合使用铅管和非铅管，33%的城市使用非铅管。由于不同城市采用的管道类型不完全相同，不同管道类型代表了铅的暴露量不同，这使分析铅的暴露量与婴儿死亡率的因果关系变得可能。基于此背景，本案例利用1900年美国172个大中型城市（这些城市包含了美国84%的人口）的相关数据，研究水源性铅暴露与婴儿死亡率之间的关系。

# 二、数据来源与变量说明

本案例使用数据集主要包含城市级别的婴儿死亡率、地理信息、人口因素、使用的供水管道类型、水质因素以及气象因素等变量数据（Clay et al., 2014），变量说明如表2.1所示。

死亡率是衡量人口健康的基本指标，高水平的铅暴露通常会造成成人发病而不是死亡，而婴儿对铅高度敏感，因此本案例选取婴儿死亡率作为衡量铅暴露对人口健康影响的指标，也就是本案例的因变量。查阅相关资料，可以使用供水管道类型（lead）衡量水源性铅暴露的状态，而管道作为输送水源的载体，其本身可能受到水质因素的影响，因此纳入水质因素相关评价变量酸碱度（pH）和硬度（hardness），特别是酸碱度（pH），酸碱度决定了水溶解管道上铅的能力，进而决定了水中铅的含量。除此之外，婴儿死亡率还可能受到本地人口因素和气象因素的影响，因此纳入人口因素变量（如城市人口population、平均年龄age、婴儿在外国出生的人口比率foreign\_share、伤寒死亡率typhoid\_rate、非肺结核病死亡率np\_tub\_rate、育龄妇女比例育龄妇女比例mom\_rate）和气象因素变量（州内平均降水量precipitation、州内平均温度temperature）。

表2.1 数据集变量说明表

|  |  |  |
| --- | --- | --- |
| **变量分类** | **变量** | **描述** |
| 因变量 | infrate | 婴儿死亡率 |
| 地理信息 | city | 城市名 |
| state | 所在州 |
| 管道类型 | lead | 二分变量。如果该城市使用含铅管道，则为1，如果没有使用含铅管道，则为0 |
| 水质因素 | pH | 水的PH值 |
| hardness | 水的硬度指数 |
| 人口因素 | population | 城市人口数量 |
| age | 平均年龄 |
| foreign\_share | 外国出生的人口比例 |
| typhoid\_rate | 伤寒死亡率 |
| np\_tub\_rate | 非肺结核病死亡率 |
| mom\_rate | 育龄妇女比例 |
| 气象因素 | precipitation | 州内平均降水量 |
| temperature | 州内平均温度 |

# 三、描述统计分析

首先，对数据进行描述性分析，检查数据质量，观察各个变量数据分布的情况，为后续的建模分析提供基础。

图3.1展示了美国172个城市的婴儿死亡率（infrate）的箱线图。可以看到，婴儿死亡率均值约为39.619%，中位数约为38.644%；最小值（下边缘）为10.976%，出现在加利福尼亚州（CA）的阿拉米达市（Alameda），最大值（上边缘）75.155%出现在缅因州（ME）波特兰市（Biddeford）。箱线图异常值（超过上四分位数1.5个四分位差）有三个，分别为79.928%、83.246%、84.447%，对应的是南卡罗来纳州（SC）查尔斯顿市（Charleston）、马萨诸塞州（MA）奇科皮市（Chicopee）和马萨诸塞州（MA）福尔里弗市（Fall River）。

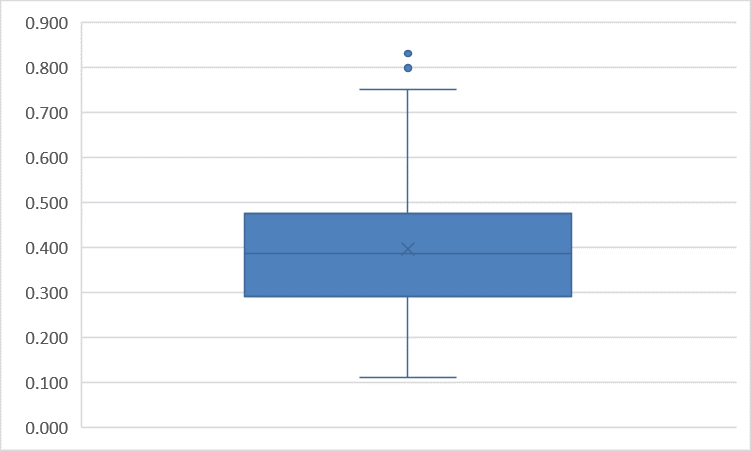


图3.1 婴儿死亡率箱线图

如下图3.2所示，在172个城市中，仅使用和混合使用含铅管道的有117个城市，占比68.023%，完全不使用含铅管道的城市有55个，占比31.977%。说明在十九世纪的美国，城市使用铅管是非常普遍的现象。

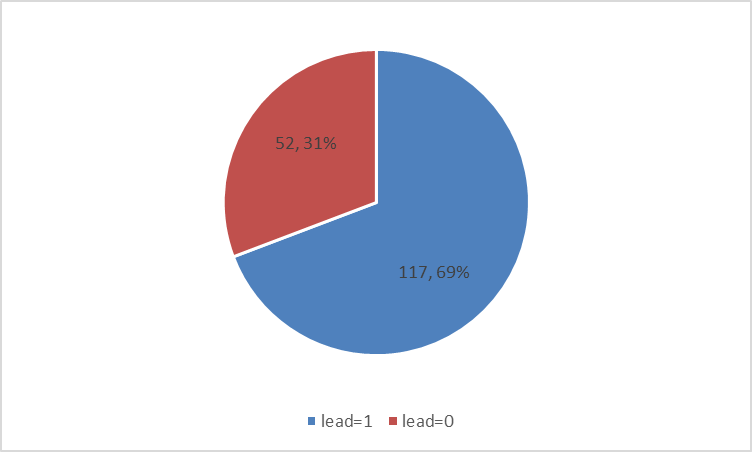


图3.2 含铅管道使用情况

表3.1、3.2、3.3分别展示了可能影响婴儿死亡率的各城市水质因素、人口因素、气象因素数据的统计量。水质因素中，代表酸碱性的变量pH的均值为7.323，中位数为7.500，可见172个城市中大部分城市的水质偏碱性；代表水质硬度的变量hardness的均值为113.436，但标准差为98.759，最大值为445.000，最小值仅为2.000，表明172个城市的水质硬度差别较大。

表3.1 水质因素描述统计表

| 统计量 | pH | hardness |
| --- | --- | --- |
| 平均 | 7.323 | 113.436 |
| 标准误差 | 0.053 | 7.530 |
| 中位数 | 7.500 | 96.500 |
| 众数 | 6.500 | 21.000 |
| 标准差 | 0.692 | 98.759 |
| 峰度 | -0.696 | 0.875 |
| 偏度 | -0.265 | 1.177 |
| 区域 | 3.200 | 443.000 |
| 最小值 | 5.700 | 2.000 |
| 最大值 | 8.900 | 445.000 |
| 观测数 | 172 | 172 |

注意：Excel描述统计功能输出的统计量有平均值、标准误差、中位数、众数、标准差、峰度、偏度、区域（即极差，代表数据最大值与最小值之差）、最小值、最大值和观测数。

表3.2显示，人口因素中城市人口（population）平均值为1087.962，平均年龄（age）的均值为27.705岁，外国人口出生比例（foreign\_share）平均值为21.7%，伤寒死亡率（typhoid\_rate）平均值为4.1%，非肺结核病死亡率（np\_tub\_rate）平均值为2%，育龄妇女占总人口比例（mom\_rate）的均值为19.9%。

表3.2 人口因素描述统计表

| 统计量 | population | age | foreign\_share | typhoid\_rate | np\_tub\_rate | mom\_rate |
| --- | --- | --- | --- | --- | --- | --- |
| 平均 | 1087.962 | 27.705 | 0.217 | 0.041 | 0.020 | 0.199 |
| 标准误差 | 239.218 | 0.169 | 0.008 | 0.002 | 0.001 | 0.002 |
| 中位数 | 357.500 | 27.589 | 0.214 | 0.034 | 0.019 | 0.197 |
| 众数 | 191.000 | 28.885 | / | 0.028 | 0.000 | 0.175 |
| 标准差 | 3137.315 | 2.217 | 0.110 | 0.029 | 0.012 | 0.032 |
| 峰度 | 78.836 | 0.126 | 0.151 | 2.562 | 0.240 | 1.099 |
| 偏度 | 8.198 | 0.237 | 0.394 | 1.589 | 0.610 | 0.540 |
| 区域 | 34270.539 | 12.431 | 0.582 | 0.144 | 0.057 | 0.193 |
| 最小值 | 85.000 | 22.013 | 0.000 | 0.000 | 0.000 | 0.131 |
| 最大值 | 34355.539 | 34.444 | 0.582 | 0.144 | 0.057 | 0.324 |
| 观测数 | 172 | 172 | 172 | 172 | 172 | 172 |

表3.3显示，气象因素中降雨量（precipitation）平均值为3.329mm，标准差为0.579mm，最大降雨量地区和最小降雨量地区相差3.799mm；各地平均温度的（temperature）平均值为49.324华氏摄氏度，最大平均温度与最小平均温度之差为29.812华氏摄氏度。

表3.3 气象因素描述统计表

|  |  |  |
| --- | --- | --- |
| 统计量 | precipitation | temperature |
| 平均 | 3.329 | 49.324 |
| 标准误差 | 0.044 | 0.389 |
| 中位数 | 3.366 | 48.524 |
| 众数 | 3.614 | 47.724 |
| 标准差 | 0.579 | 5.096 |
| 峰度 | 2.494 | 2.326 |
| 偏度 | -1.221 | 1.239 |
| 区域 | 3.799 | 29.812 |
| 最小值 | 0.955 | 40.811 |
| 最大值 | 4.754 | 70.623 |
| 观测数 | 172 | 172 |

# 四、假设检验

### 4.1城市人口数量和含铅管道使用的独立性检验

为什么有的城市使用含铅管道，而有的城市不使用？这是一个完全随机的现象吗？还是背后另有原因？通过查阅资料，发现人口总量越大的城市越倾向于使用含铅管道，人口的数量可能和含铅管道使用有关。因此，我们设置了如下列联表独立性检验：假设人口总数小于21,500人的城市为少人口城市，大于80,400人的城市为多人口城市，介于两者之间的城市为中等人口城市，于是我们可以得到城市人口水平（少人口、中等人口、多人口）与是否使用含铅管道（lead=0、lead=1）之间的独立性假设检验问题。下表4.1城市人口水平与是否使用含铅管道（lead）的列联表。

表4.1 城市人口水平与水管类别列联表

|  |  |  |  |
| --- | --- | --- | --- |
|  | 铅管 | 非铅管 | 总和 |
| 少人口 | 25 | 18 | 43 |
| 中等人口 | 53 | 32 | 85 |
| 多人口 | 39 | 5 | 44 |
| 总和 | 117 | 55 | 172 |

进一步计算得到列联表各单元的频数期望值，计算结果如下表4.2所示：

表4.2 列联表的期望值计算结果

|  |  |  |  |
| --- | --- | --- | --- |
|  | 铅管 | 非铅管 | 总和 |
| 少人口 | 29.250 | 13.750 | 43 |
| 中等人口 | 57.820 | 27.180 | 85 |
| 多人口 | 29.930 | 14.070 | 44 |
| 总和 | 117 | 55 | 172 |

由表4.1和4.2可以计算检验统计量为：

此处城市人口水平有3个，是否使用含铅管道（lead）水平有2个，故自由度为。

故可以拒绝原假设，认为城市人口水平与是否使用含铅管道不独立，因此可以证明城市人口数量确实和含铅管道的使用有关。

### 4.2方差分析：含铅管道的使用对婴儿死亡率是否有显著影响？

既然铅的暴露量可由是否使用含铅管道，即lead变量表示，那么在使用含铅管道和不使用含铅管道的城市里，婴儿的死亡率是否有显著的差异呢？

本案例数据集中变量lead是0-1变量，lead因子有两个水平，分别为lead=1表示仅使用和混合使用含铅管道，lead=0表示不使用含铅管道。对于这种数据，我们可以考虑进行单因素方差分析来验证lead变量处于不同水平时，婴儿死亡率的均值是否不一样。即作如下假设检验：

在进行方差分析前，还需要检验问题是否满足以下假设：

①每一总体是否为正态分布(绘制直方图观察，更严谨一些可以使用正态性检验)；

②每一总体的方差是否相等（方差齐性检验，在这里可以使用F检验）；

③每一总体中的样本结果相互独立。

在本例中，我们绘制直方图观察，数据大致呈现出了正态分布先上升再下降的对称趋势（图4.1、图4.2）。

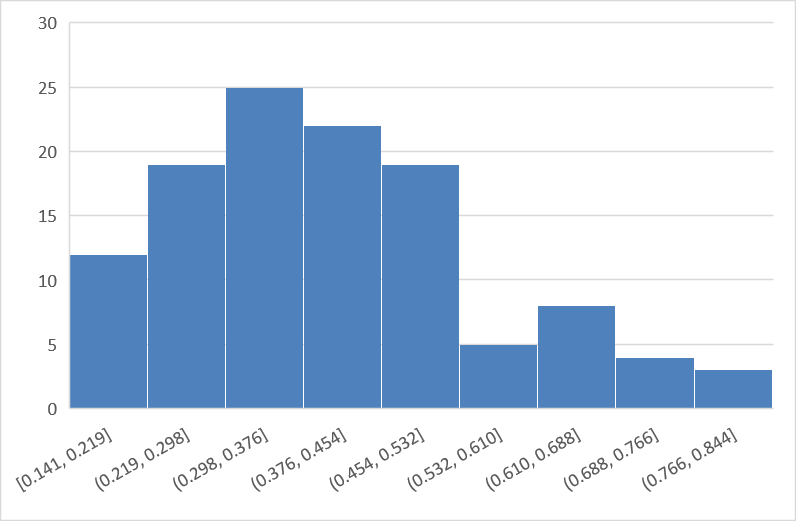


图4.1 lead=1直方图

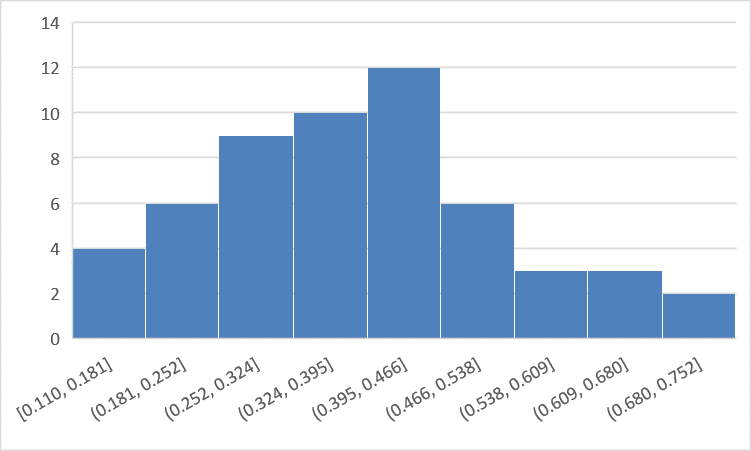


图4.2 lead=0直方图

由于问题中只有两个水平，故可以使用F检验来判断方差是否相等，在Excel的数据分析中选择F检验-双样本方差得到如下结果：

表4.3 F-检验 双样本方差分析表

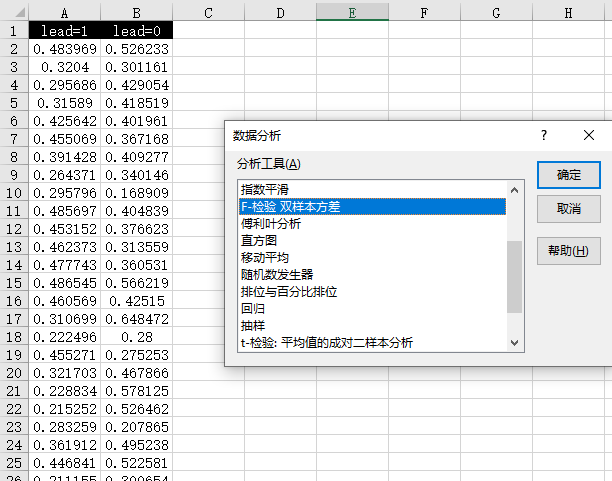
|  |  |  |
| --- | --- | --- |
|  | lead=1 | lead=0 |
| 平均 | 0.403 | 0.381 |
| 方差 | 0.023 | 0.022 |
| 观测值 | 117 | 55 |
| df | 116 | 54 |
| F | 1.073 |  |
| P(F<=f) 单尾 | 0.392 |  |
| F 单尾临界 | 1.494 |  |

可以看到，，不能拒绝方差相等的原假设，故认为该问题可以使用单因素方差分析检验。

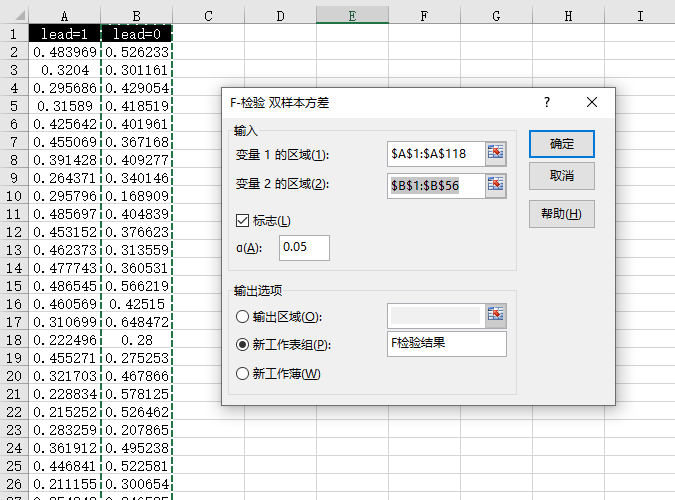
#### *用Excel进行F检验双样本方差分析的操作步骤*

**Step 1：**点击【数据】，并点击【数据分析】选项

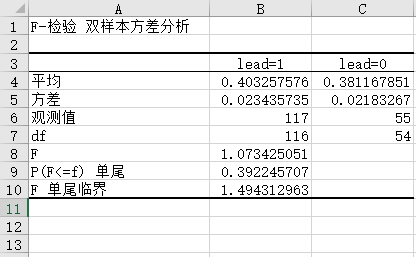
**Step 2：**在分析工具中选择【F检验：双样本方差】，点击【确定】



**Step 3：**当对话框出现时：勾选【标志】，在【变量1的区域】选择单元格区域A1:A118，在【变量2的区域】选择单元格区域B1:B56，【】方框内根据需要填入相应的值，这里我们选择0.05，在【输出选项】中根据需要选择输出位置，这里我们选择新工作表组并给定命名“F检验结果”，点击确定后，得到输出结果。



**Step 4：**根据输出结果进行分析。



确定该问题可以使用单因素方差分析后，使用Excel数据分析中的单因素方差分析功能，得到结果如下表4.4所示，，不能拒绝不同水平均值相等的原假设。因此，可认为在使用含铅管道和不使用含铅管道的城市里，婴儿的死亡率没有显著差异。

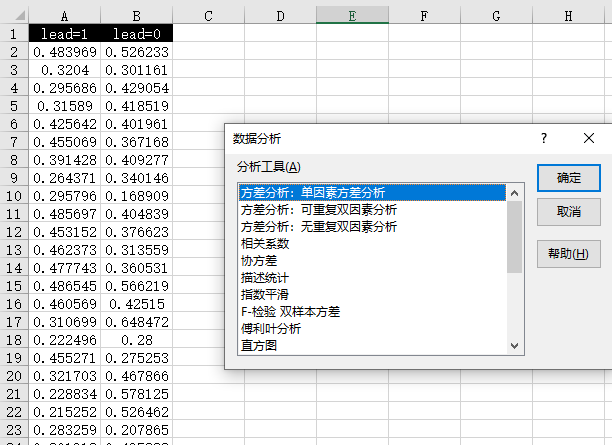
表4.4 方差分析：单因素方差分析

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| SUMMARY |  |  |  |  |  |  |
| 组 | 观测数 | 求和 | 平均 | 方差 |  |  |
| lead=1 | 117 | 47.181 | 0.403 | 0.023 |  |  |
| lead=0 | 55 | 20.964 | 0.381 | 0.022 |  |  |
| 方差分析 |  |  |  |  |  |  |
| 差异源 | SS | df | MS | F | P-value | F crit |
| 组间 | 0.018 | 1 | 0.018 | 0.796 | 0.373 | 3.897 |
| 组内 | 3.898 | 170 | 0.023 |  |  |  |
| 总计 | 3.916 | 171 |  |  |  |  |

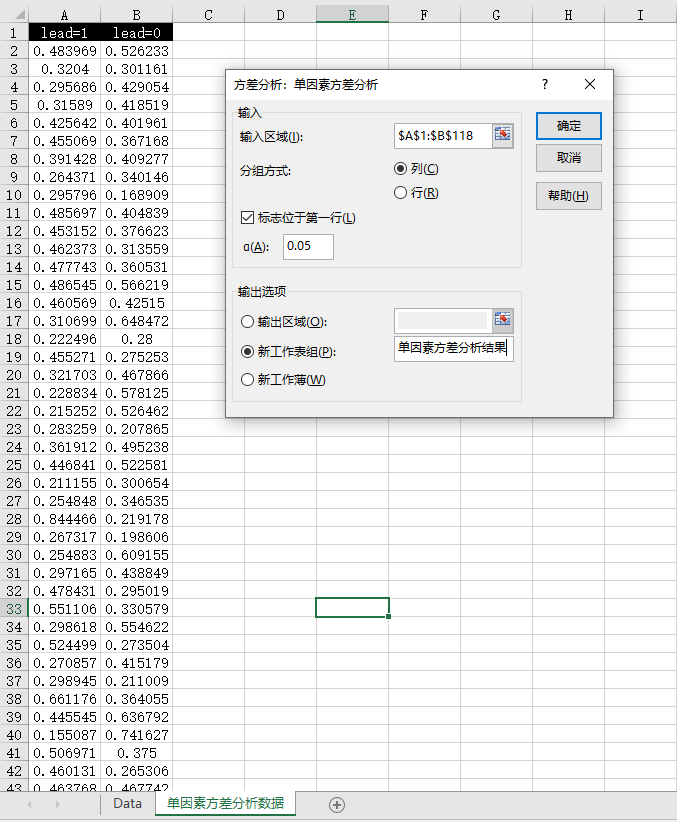
#### *用Excel进行单因素方差分析的操作步骤*

**Step 1：**点击【数据】，并点击【数据分析】选项

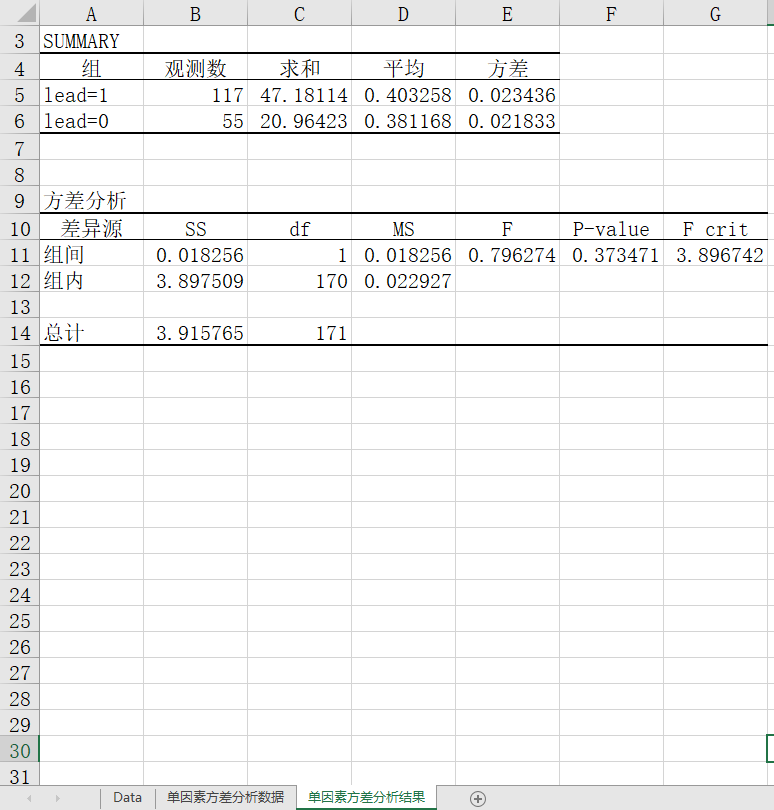
**Step 2：**在分析工具中选择【方差分析：单因素方差分析】，点击【确定】



**Step 3：**当对话框出现时：勾选【标志位于第一行】，在【输入区域】部分选择数据单元格区域A1:B118，【】方框内根据需要填入相应的值，这里我们选择0.05，在【输出选项】中根据需要选择输出位置，这里我们选择新工作表组并给定命名“单因素方差分析结果”，点击确定后，得到输出结果。



**Step 4：**根据输出结果进行分析。



### 4.3大样本Z检验：非正态总体的假设检验

在正态、方差齐性、样本独立的假设下，我们在上一小节进行了单因素方差分析，得到了lead因子不显著的结果。但是单因素方差分析的前提假设过于严格（在上述显示的lead不同水平下的婴儿死亡率分布的直方图（图4.1、4.2）并非严格服从正态分布），能否采用放宽条件的假设检验方法?

我们知道，在两样本均值差异检验问题中，如果两样本为独立样本，并且都为大样本()，即使两总体不服从正态分布，也可以使用正态Z检验。因此我们使用均值正态Z检验进一步验证我们的假设。在Excel的数据分析中选择Z检验：双样本平均差检验，给定假设平均值为0，变量lead=1的方差为其自身的方差估计0.023，变量lead=0的方差为其自身的方差估计0.022，得到结果如下表4.5所示，，同样说明了不同lead水平下的婴儿死亡率均值差异不显著。

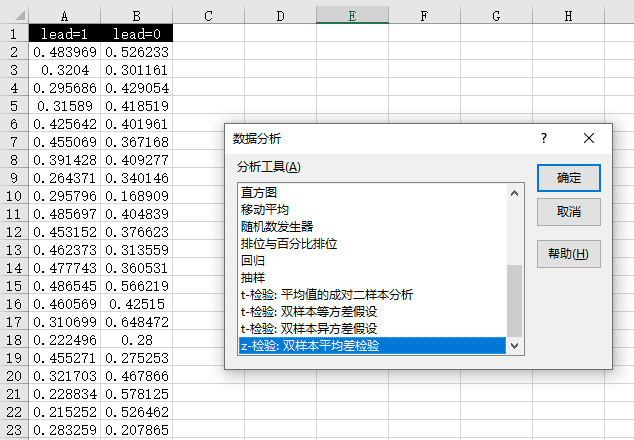
表4.5 z-检验: 双样本均值差异分析

|  |  |  |
| --- | --- | --- |
|  | lead=1 | lead=0 |
| 平均 | 0.403 | 0.381 |
| 已知协方差 | 0.023 | 0.022 |
| 观测值 | 117 | 55 |
| 假设平均差 | 0 |  |
| z | 0.904 |  |
| P(Z<=z) 单尾 | 0.183 |  |
| z 单尾临界 | 1.645 |  |
| P(Z<=z) 双尾 | 0.366 |  |
| z 双尾临界 | 1.960 |  |

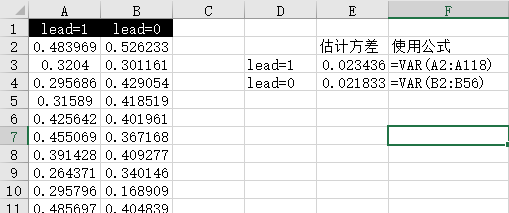
#### *用Excel进行Z检验的操作步骤*

**Step 1：**点击【数据】，并点击【数据分析】选项

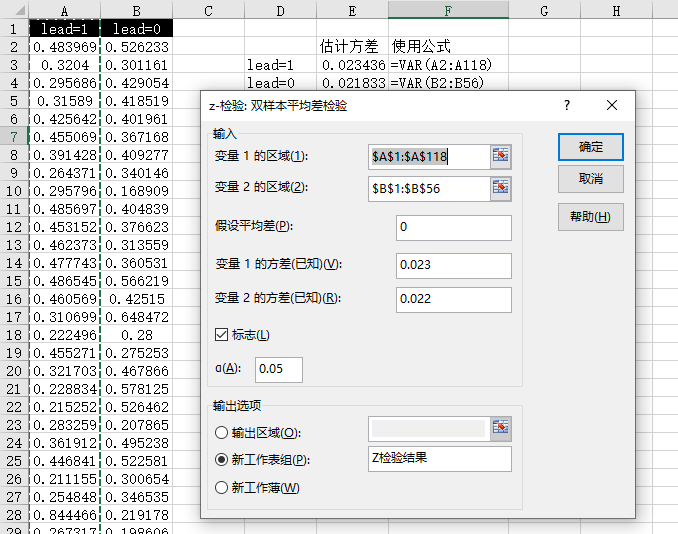
**Step 2：**在分析工具中选择【Z检验：双样本平均值检验】，点击【确定】



**Step 3：**当对话框出现时：勾选【标志】，在【变量1的区域】选择单元格区域A1:A118，在【变量2的区域】选择单元格区域B1:B56，【假设平均值】输入0，【变量1的方差】输入估计值0.023，【变量2的方差】输入估计值0.022。



【】方框内根据需要填入相应的值，这里我们选择0.05，在【输出选项】中根据需要选择输出位置，这里我们选择新工作表组并给定命名“Z检验结果”，点击确定后，得到输出结果。



**Step 4：**根据输出结果进行分析。



# 五、回归分析

上述4.2节和4.3节中的两种检验方式都不能证明在使用铅管的城市里婴儿的死亡率更高，这是否就说明婴儿的死亡率和铅管的使用无关呢？其实不然，由于一个城市的婴儿死亡率受非常多变量的影响，婴儿的死亡率和铅管的使用之间的关系可能会受到其它变量的干扰。举例而言，我们已经发现城市越大越有可能使用铅管，使用铅管在理论上可能会提高婴儿死亡率，但是大城市可能有更好的水质，更好的水质会有更低的婴儿死亡率，两者作用相互抵消，导致最后使用铅管的城市的婴儿死亡率并不比不使用铅管的城市高。因此，仅用方差分析或者大样本两均值差异检验并不能证明婴儿的死亡率和铅管的使用是完全无关的，我们需要引入新的变量（本质上就是控制其它因素），通过回归建模分析得到它们之间的关系。

### 5.1考虑水质酸碱性的回归建模

城市中的水资源是所有人生活的必需品，水的质量在极大程度上影响人们的健康，世界历史上也多次出现水污染而造成城市大面积发病的现象。而衡量水质的一个重要因素就是pH值，因此，我们可以考虑在婴儿死亡率和铅管使用的相关性分析中引入pH值变量，建立以下回归模型：

这里我们对pH值做了变换，用pH减去5.675。原因是6.675是pH值的0.25分位点，我们希望在0.25分位点时，它的值为0，所以我们进行了对数处理。结果如下表5.1所示。

表5.1 回归分析结果



回归结果显示使用铅管会提高婴儿死亡率，其p值为0.097，说明此种影响较为显著。而pH值则对婴儿死亡率有非常显著的影响，pH值升高将会显著降低婴儿死亡率，说明水的酸性越高，对于婴儿的成长越不利。

### 5.2引入交叉项的回归建模

pH值除了可能会直接影响婴儿死亡率外，其与含铅管道使用的交互作用也不能被忽视。其原因是在使用铅管的城市里，水能够溶解多少铅管壁上的铅和水的pH值有关，如果水的酸性越高（pH值越低），那么溶解的铅也就越多。按照上述的理论，我们可以添加交互项，得到以下回归方程。

回归中的第三项是一个交叉项，代表了是否使用含铅管道（Lead）与pH值的交互作用。交叉项前的系数该如何解释？它代表了相比于没有使用铅管道的城市，在使用铅管道的城市中，pH值对婴儿死亡率造成的额外的影响。

用Excel数据分析中的回归模型，得到以下结果。

表5.2 引入交叉项回归分析结果



回归系数，说明当pH值为6.675时，使用铅管将使婴儿死亡率提高4.8%，而由于，说明这个影响在5%的置信水平下是显著的。在5%的置信水平下，其置信区间为。

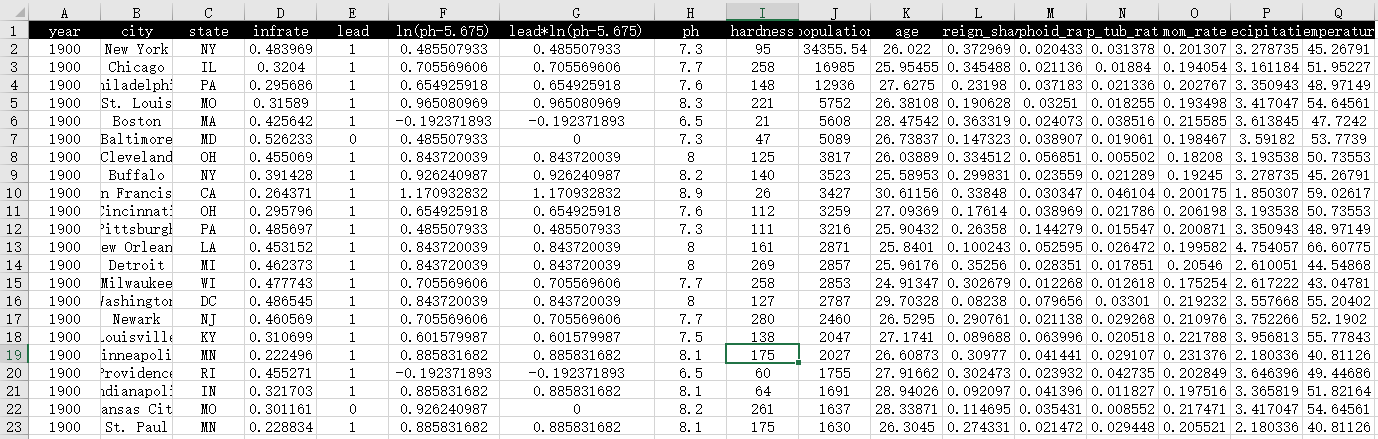
回归系数，该系数的含义为相比于没有使用铅管的城市，（pH-5.675）增加1%，会多使婴儿死亡率降低0.043%，但是该系数的p值大于0.1，说明此种影响并不显著。

最后进行线性关系检验，检验因变量与自变量之间的线性关系是否显著，即模型总体显著性检验。结果表明，因此拒绝原假设，表明其线性关系显著。

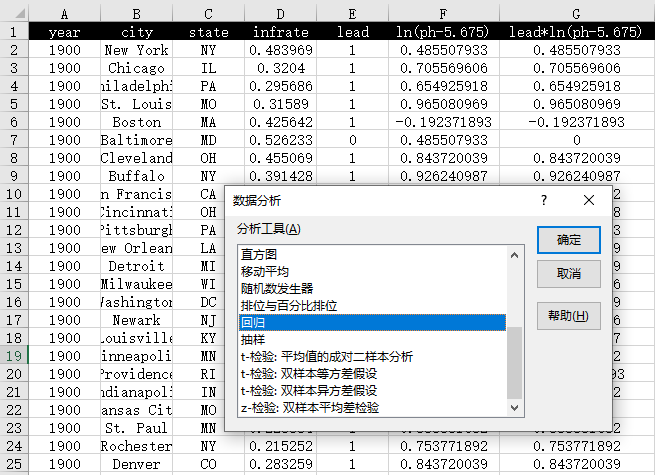
总体而言，回归结果显示城市使用铅管情况与水的酸碱度对于该城市婴儿的死亡率都有显著的影响，但是两者的交叉影响并不显著。回归的拟合优度为0.231，表明该回归方程所能解释的因变量变异性的百分比为23.1%。

#### *用Excel进行回归分析的步骤*

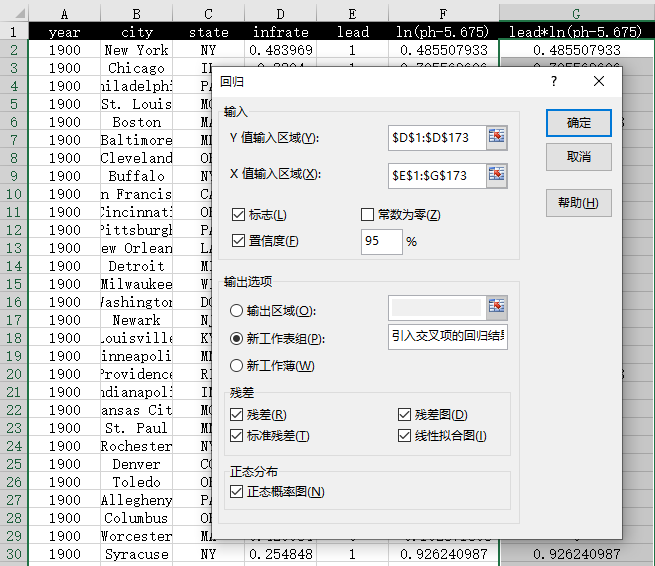
**Step 1：**将要回归的数据处理好，这里注意要纳入模型的自变量列数需连续放置。点击【数据】，并点击【数据分析】选项



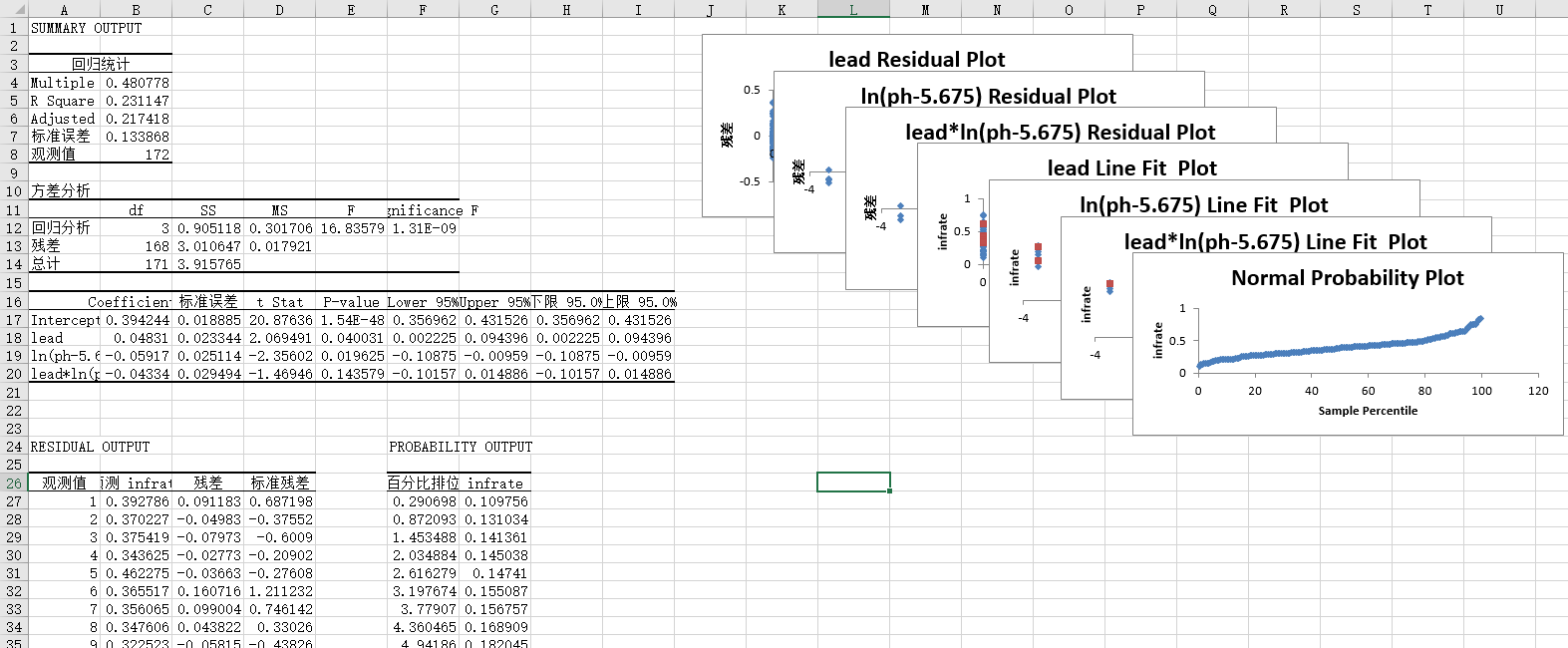
**Step 2：**在分析工具中选择【回归】，点击【确定】



**Step 3：**当对话框出现时：勾选【标志】，在【Y值输入区域】选择数据单元格区域D1:D173，在【X值输入区域】选择数据单元格区域E1:G173，【置信度】方框内根据需要填入相应的值，这里我们输入95，在【输出选项】中根据需要选择输出位置，这里我们选择新工作表组并给定命名“引入交叉项的回归结果”，【残差】勾选需要输出的残差检验相关信息，这里我们全部勾选，【正态分布】勾选输出正态概率图，点击确定后，得到输出结果。



**Step 4：**根据输出结果进行分析。



### 5.3多元线性回归建模

然而，5.2节回归的结果显示为0.231，显然解释的力度还不够，因此，我们尝试在5.2节的基础上继续引入其它变量。在5.2节的基础上进一步加入其它的控制变量（除地理因素外的其它所有变量），用婴儿死亡率对所有的变量做回归，回归方程为：

其中，新加的***X***表示所有的控制变量，包括水的硬度（取对数）、人口（取对数）、城市平均年龄、外来出生的人口的比例、伤寒死亡率、非肺结核死亡率、孕龄妇女比例、州内平均降水量和州内温度。

进行多元回归之前，首先要进行多重共线性的检测，由于变量较多，可以利用相关系数矩阵来判断。

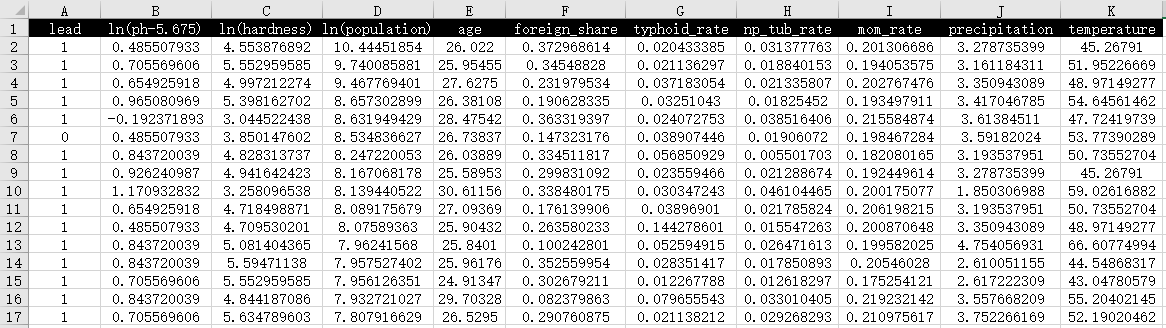
表5.3 相关系数矩阵



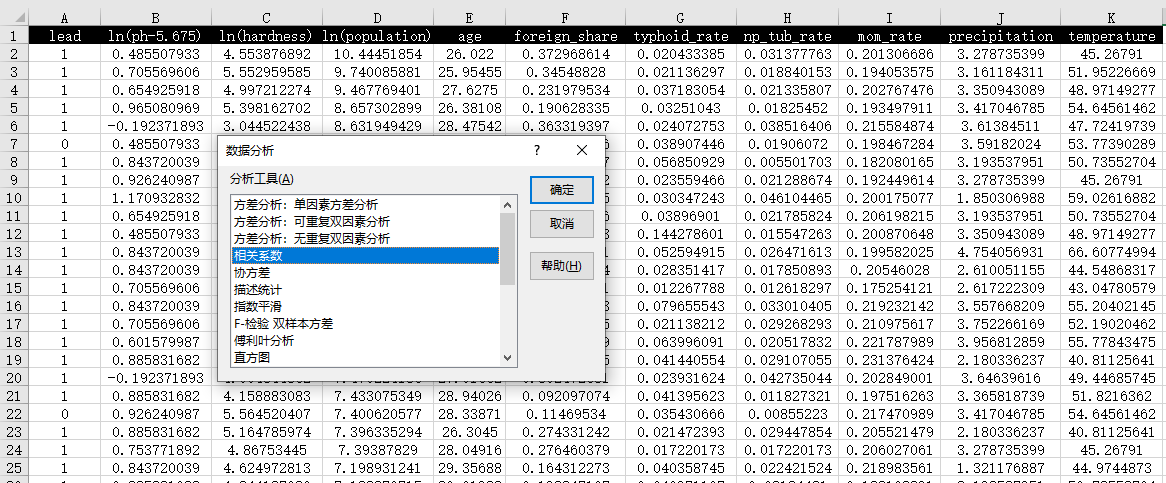
可以观察到除了pH和水体硬度之间的相关性达到0.733，其余的变量之间的相关性（绝对值）不超过0.43，因此可初步认为其它变量之间没有多重共线性，而pH值和水体硬度之间则再做进一步的检测。

#### *用Excel计算相关系数矩阵的步骤*

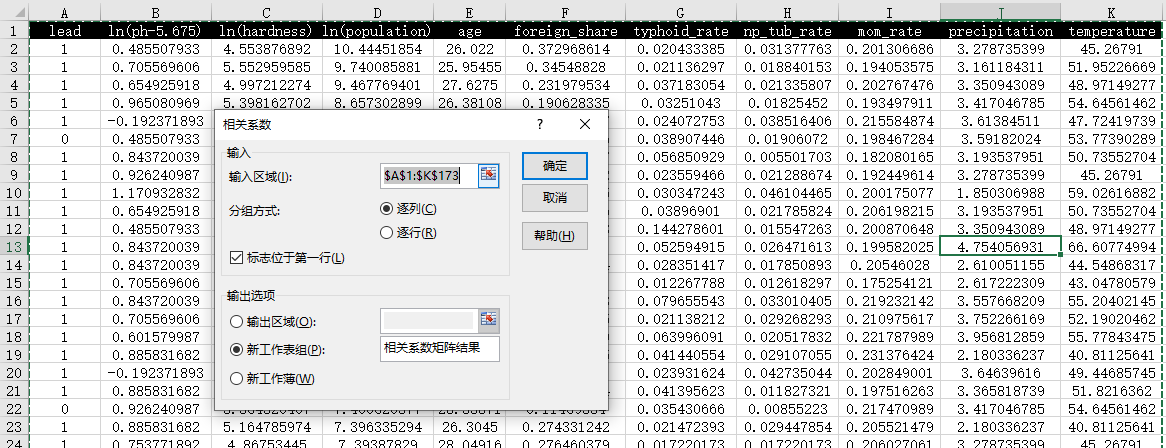
**Step 1：**将要计算相关系数矩阵的数据处理好，这里注意参与计算的自变量列数需连续放置。点击【数据】，并点击【数据分析】选项



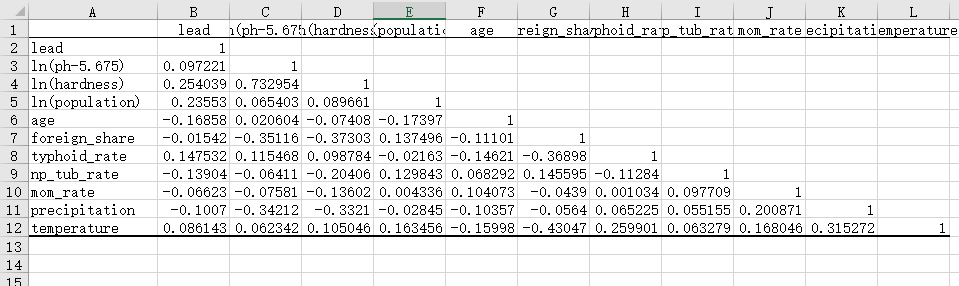
**Step 2：**在分析工具中选择【相关系数】，点击【确定】



**Step 3：**当对话框出现时：勾选【标志位于第一行】，在【输入区域】选择数据单元格区域A1:K173，在【输出选项】中根据需要选择输出位置，这里我们选择新工作表组并给定命名“相关系数矩阵结果”，点击确定后，得到输出结果。



**Step 4：**根据输出结果进行分析。



容忍度与方差扩大因子（VIF）通常作为衡量多重共线性的方法。容忍度越小，多重共线性越严重。通常认为容忍度小于0.1时，存在严重的多重共线性。方差扩大因子（VIF）等于容忍度的倒数，显然，VIF越大，多重共线性越严重。一般认为VIF大于10时，存在多重共线性问题。这里我们利用VIF来检测是否存在多重共线性。

建立ln(pH-5.675)与其他自变量之间的回归模型，结果如表5.4所示，计算得，则，故可以认为该问题中的自变量不存在多重共线性问题。

表5.4 回归分析结果



检查多重共线性的问题之后，利用Excel的数据分析中的回归模型进行回归，结果如下图所示。

表5.5 回归分析结果



回归结果显示，为0.568，调整后的为0.535，相比5.2节中的回归的有显著的提升，说明其它的变量对于因变量也有很强的解释能力。与5.2节中的回归不同的是，pH值对于婴儿死亡率的影响变得不再显著，但是pH和铅管使用的交互变量变得显著，说明了在使用铅管的城市里，pH值对婴儿死亡率有影响，因为更低的pH值导致吸收了更多的铅，而在不使用铅管的城市中，pH值对于婴儿死亡率没有显著影响。

在新的回归结果下，当pH值为6.675时，使用铅管将使得婴儿死亡率提高4.3%。而在控制其它变量不变的条件下，降雨量的提高伤寒病死率的提高都会使婴儿的死亡率提高。

在得到回归结果后，还需对模型进行诊断。由于回归模型需要满足一定的假设，因此需要保证假设都能够满足，我们通常的做法是绘制残差图，以保证方差齐性，这里具体的残差图如下（利用回归模型输出的预测值与标准残差绘制散点图）。

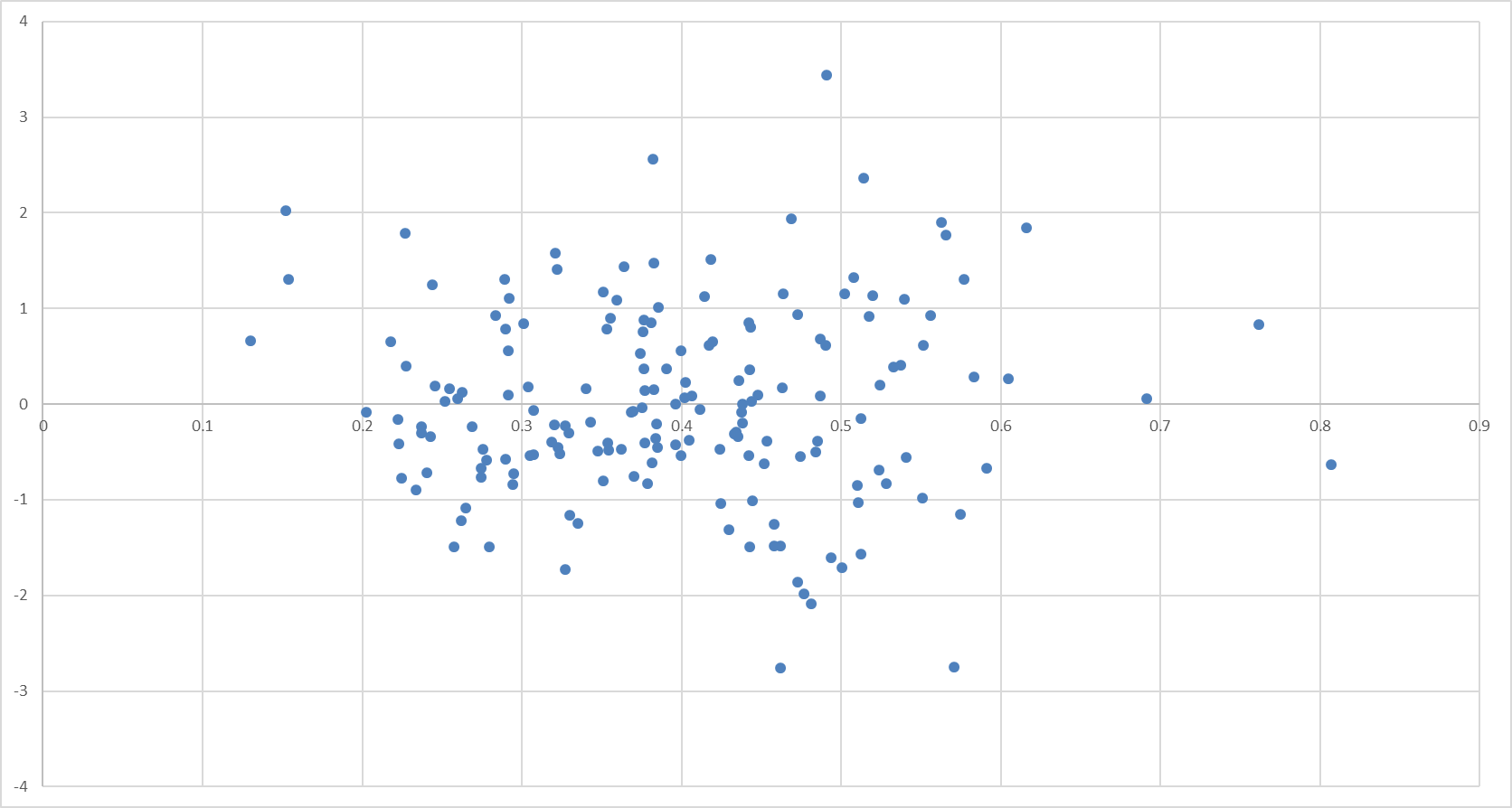


图5.1 残差图（预测值与标准残差）

图中没有出现明显的方差随着预测值的变化而变化的特点，说明该回归模型能够满足方差齐性。

最后，如果想要用尽可能少的变量得到回归结果，可以用向后剔除法得到需要的变量，具体流程如下：

首先观察两个最不显著的变量，人口和孕龄妇女的比例。去掉使得SSE值减小的最小的变量，即孕龄妇女比例指标。去掉孕龄妇女指标之后，再去掉人口指标。（虽然pH的系数也不显著，但是和lead的交叉项显著，因此不能去掉），比较温度和水的硬度指标，发现温度使回归SSE减小的更少，去掉温度指标。去掉水的硬度指标，如表5.6所示，最后所有的检验统计量都显著。

表5.6 回归分析结果



# 延伸阅读

Clay, K., Troesken, W., Haines, M.. (2014). Lead and mortality. NBER Working Papers, 96(3), 458-470.