窗体顶端

|  |  |
| --- | --- |
|  | 第6章课后习题，第4题：多元线性回归分析 |
|  | 假设8个样例的数据集如下，其中和是解释变量，是被解释变量。借助软件（如SPSS或C++配套库）进行线性回归分析。

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 1.9 | 2.9 | 2.1 | 3.6 | 4.5 | 3.8 | 4.7 | 4.1 |
|  | 8.2 | 8.3 | 8.7 | 9.7 | 9.3 | 9.6 | 9.2 | 9.6 |
|  | 3.9 | 4.8 | 4.8 | 4.6 | 4.7 | 4.3 | 4.5 | 3.8 |
|  | 9.1 | 9.6 | 10.5 | 11 | 10.8 | 11.2 | 10.6 | 9.7 |

（1）请计算各变量之间的相关系数，并绘制散点图。（2）建立多元线性回归方程。（3）进行参数检验，按显著性水平0.05，判别参数的显著性。（4）如果存在不显著的解释变量，分析其原因。本题中采用剔除后再建模。（5）进行回归模型检验，并根据判定系数和修正判定系数分析模型拟合精度。（6）进行残差分析，判别异常样例并进行DW检验。（7）请依据回归结果对回归系数进行解释。（8）给定新样例“”，预测值，给出置信度95%的预测区间。 |
| Spss | 1. 双变量相关性分析：

点击“分析”栏中“相关”选项下“双变量”，对多个变量之间做两两相关性分析：散点图绘制：点击图形下栏“图形画板模板选择器”，shift选中所有变量，点击“散点图矩阵即可”，取最下排一行散点图即可。（2）点击“分析”栏中“回归”选项下“线性”，做多元线性回归：（3）操作同（2）。（4）剔除不显著的变量后，对其余变量同（2）操作。（5）同（4）（6）DW检验：在线性回归右上角“统计”选项中点选“德宾沃森”。残差分析：在线性回归中，“图”选项中使y轴为ZRESID,x轴为ZPRED（7）略（8）通过（5）中建立的线性回归方程，计算出y值，再进行线性回归，点选“保存”选张中预测区间“单值”，如下： |
| C++ |  |
| 输出 | （1）相关性分析结果如下

|  |
| --- |
| **相关性** |
|  | **x1** | **x2** | **x3** | **y** |
| **x1** | **皮尔逊相关性** | **1** | **0.747\*** | **0.021** | **0.523** |
| **Sig.（双尾）** |  | **0.033** | **0.961** | **0.183** |
| **个案数** | **8** | **8** | **8** | **8** |
| **x2** | **皮尔逊相关性** | **0.747\*** | **1** | **-0.113** | **0.716\*** |
| **Sig.（双尾）** | **0.033** |  | **0.791** | **0.046** |
| **个案数** | **8** | **8** | **8** | **8** |
| **x3** | **皮尔逊相关性** | **0.021** | **-.113** | **1** | **0.478** |
| **Sig.（双尾）** | **0.961** | **.791** |  | **0.231** |
| **个案数** | **8** | **8** | **8** | **8** |
| **y** | **皮尔逊相关性** | **0.523** | **0.716\*** | **0.478** | **1** |
| **Sig.（双尾）** | **0.183** | **0.046** | **0.231** |  |
| **个案数** | **8** | **8** | **8** | **8** |
| **\*. 在 0.05 级别（双尾），相关性显著。** |

x1、x2之间相关性强，x3与x1、x2相关性不显著。散点图如下：（2）多元线性回归结果如下：

|  |
| --- |
| **系数a** |
| **模型** | **未标准化系数** | **标准化系数** | **t** | **显著性** |
| **B** | **标准错误** | **Beta** |
| **1** | **(常量)** | **-4.552** | **3.703** |  | **-1.229** | **.286** |
| **x1** | **-0.117** | **0.218** | **-0.164** | **-0.540** | **0.618** |
| **x2** | **1.136** | **0.384** | **0.904** | **2.956** | **0.042** |
| **x3** | **1.121** | **0.391** | **0.583** | **2.866** | **0.046** |
| **a. 因变量：y** |

得到多元线性回归方程应为：y=-0.117x1+1.136x2+1.121x3-4.552.（3）由（2）中结果可知，x1的系数不显著，x2和x3的系数均显著。（4）根据（3）中分析结果和多元线性分析同时输出的F检验结果：

|  |
| --- |
| **ANOVAa** |
| **模型** | **平方和** | **自由度** | **均方** | **F** | **显著性** |
| **1** | **回归** | **3.336** | **3** | **1.112** | **7.030** | **.045b** |
| **残差** | **.633** | **4** | **.158** |  |  |
| **总计** | **3.969** | **7** |  |  |  |
| **a. 因变量：y** |
| **b. 预测变量：(常量), x3, x1, x2** |

，x1对应系数不显著，整个回归方程显著，剔除不显著的x1解释变量，重新进行多元线性回归分析，结果如下：

|  |
| --- |
| **系数a** |
| **模型** | **未标准化系数** | **标准化系数** | **t** | **显著性** |
| **B** | **标准错误** | **Beta** |
| **1** | **(常量)** | **-3.388** | **2.789** |  | **-1.215** | **0.279** |
| **x2** | **0.980** | **0.234** | **0.780** | **4.189** | **0.009** |
| **x3** | **1.087** | **0.358** | **0.566** | **3.040** | **0.029** |
| **a. 因变量：y** |

得到多元线性回归方程：y=0.980x2+1.087x3-3.388.（5）根据（4）中分析结果和多元线性分析同时输出的F检验结果：

|  |
| --- |
| **ANOVAa** |
| **模型** | **平方和** | **自由度** | **均方** | **F** | **显著性** |
| **1** | **回归** | **3.290** | **2** | **1.645** | **12.115** | **0.012b** |
| **残差** | **0.679** | **5** | **0.136** |  |  |
| **总计** | **3.969** | **7** |  |  |  |
| **a. 因变量：y** |
| **b. 预测变量：(常量), x3, x2** |

从F检验结果知，回归模型显著。

|  |
| --- |
| **模型摘要b** |
| **模型** | **R** | **R 方** | **调整后 R 方** | **标准估算的错误** | **德宾-沃森** |
| **1** | **0.910a** | **0.829** | **0.761** | **0.36848** | **2.094** |
| **a. 预测变量：(常量), x3, x2** |
| **b. 因变量：y** |

其中，R方为0.829，调整R方为0.761，偏大，说明模型拟合效果很好。 （6）如上图中德宾沃森系数为2.094，接近2，故不存在自相关性。残差图：可以看到，残差上下分布均匀，且不超过2个单位，无异常值。（7）根据y=0.980x2+1.087x3-3.388可知，y随着x2、x3的增大而增大，且增幅均接近1.（8）y=10.24，其预测区间为[9.35138, 11.11736]. |
| 书籍 | 姜维.《数据分析与数据挖掘》.电子工业出版社. 2023. |
| 软件 | Spss 26 |

窗体底端