**回归方程模型的构建及变量的预测**

针对数据MRA\_B，我们将探索哪些自变量（X1, X2, X3, X4, X5, X6）在预测因变量Y的变化中是有效的，并构建相应的回归方程模型，对回归方程的有效性、自变量对因变量的预测作用进行评估。有时在相关系数显著的情况下，一些变量仍然不能够预测Y变量的变化，对此我们将试着进行解释。

【分析】

我们将分别使用forward和backward方法对回归方程模型进行选择。

【forward方法】

**表1. 皮尔逊相关系数分析表**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Y | | X1 | X2 | X3 | X4 | X5 | X6 |
| Y | | - |  |  |  |  |  |  |
| X1 | | .825\*\* | - |  |  |  |  |  |
| X2 | | .426\* | .558\*\* | - |  |  |  |  |
| X3 | | .624\*\* | .597\*\* | .493\*\* | - |  |  |  |
| X4 | | .590\*\* | .669\*\* | .445\* | .640\*\* | - |  |  |
| X5 | | 0.156 | 0.188 | 0.147 | 0.116 | .377\* | - |  |
| X6 | | 0.155 | 0.225 | 0.343 | .532\*\* | .574\*\* | 0.283 | - |

因为X1与Y的相关性最大，所以X1是最早被加工模型的变量。

在控制X1的情况下，剩下所有的变量对Y的p-enter>0.05，又X1对Y的回归系数显著，所以回归模型最终选择的变量只有X1。

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **表2. 排除的变量a** | | | | | | |
| 模型 | | 输入 Beta | t | 显著性 | 偏相关 | 共线性统计 |
| 容差 |
| 1 | X2 | -.050b | -0.386 | 0.702 | -0.074 | 0.688 |
| X3 | .204b | 1.571 | 0.128 | 0.289 | 0.644 |
| X4 | .068b | 0.470 | 0.642 | 0.090 | 0.552 |
| X5 | .002b | 0.014 | 0.989 | 0.003 | 0.965 |
| X6 | -.032b | -0.287 | 0.777 | -0.055 | 0.950 |
| a. 因变量：Y | | | | | | |
| b. 模型中的预测变量：(常量), X1 | | | | | | |

|  |  |  |  |
| --- | --- | --- | --- |
| **表3. 输入/除去的变量a** | | | |
| 模型 | 输入的变量 | 除去的变量 | 方法 |
| 1 | X1 |  | 向前（准则：要输入的 F 的概率 <= .050） |
| a. 因变量：Y | | | |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **表4. 系数a** | | | | | | |
| 模型 | | 未标准化系数 | | 标准化 | t | 显著性 |
| B | 标准错误 | Beta |
| 1 | (常量) | 14.376 | 6.620 |  | 2.172 | 0.039 |
| X1 | 0.755 | 0.098 | 0.825 | 7.737 | 0.000 |
| a. 因变量：Y | | | | | | |

回归方程为：Y = 14.376 + 0.755\*X1

ANOVA分析知F(1,28) = 59.861, *p*<0.001, 即该回归方程模型有效。自变量对因变量的决定系数R2=0.681，调整后的R2=0.670，即Y的变异中有68.1%可由X1解释。

【backward方法】

**表1. 皮尔逊相关系数分析表**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Y | | X1 | X2 | X3 | X4 | X5 | X6 |
| Y | | - |  |  |  |  |  |  |
| X1 | | .825\*\* | - |  |  |  |  |  |
| X2 | | .426\* | .558\*\* | - |  |  |  |  |
| X3 | | .624\*\* | .597\*\* | .493\*\* | - |  |  |  |
| X4 | | .590\*\* | .669\*\* | .445\* | .640\*\* | - |  |  |
| X5 | | 0.156 | 0.188 | 0.147 | 0.116 | .377\* | - |  |
| X6 | | 0.155 | 0.225 | 0.343 | .532\*\* | .574\*\* | 0.283 | - |

模型1含所有的预测变量；因为X5变量的预测系数В=0.031，p-remove=0.796，为其中的最大值，所以模型2剔除X5变量；同理，X4、X2、X6、X3变量对应的p-remove>0.1, 所以依次剔除。最后回归模型6中的预测变量只剩X1，X1对Y的回归系数显著，p-remove<0.1。

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **排除的变量a** | | | | | | | |
| 模型 | | 输入 Beta | | t | 显著性 | 偏相关 |  |
| 容差 |
| 2 | X5 | .031b | | 0.261 | 0.796 | 0.054 | 0.814 |
| 3 | X5 | .044c | | 0.387 | 0.702 | 0.079 | 0.884 |
| X4 | .084c | | 0.470 | 0.643 | 0.095 | 0.353 |
| 4 | X5 | .043d | | 0.391 | 0.699 | 0.078 | 0.884 |
| X4 | .091d | | 0.518 | 0.609 | 0.103 | 0.354 |
| X2 | -.077d | | -0.588 | 0.562 | -0.117 | 0.628 |
| 5 | X5 | .001e | | 0.007 | 0.995 | 0.001 | 0.965 |
| X4 | -.024e | | -0.157 | 0.876 | -0.031 | 0.462 |
| X2 | -.104e | | -0.799 | 0.432 | -0.155 | 0.648 |
| X6 | -.158e | | -1.291 | 0.208 | -0.245 | 0.704 |
| 6 | X5 | .002f | | 0.014 | 0.989 | 0.003 | 0.965 |
| X4 | .068f | | 0.470 | 0.642 | 0.090 | 0.552 |
| X2 | -.050f | | -0.386 | 0.702 | -0.074 | 0.688 |
| X6 | -.032f | | -0.287 | 0.777 | -0.055 | 0.950 |
| X3 | .204f | | 1.571 | 0.128 | 0.289 | 0.644 |
| a. 因变量：Y | | | | | | | |
| b. 模型中的预测变量：(常量), X6, X1, X2, X3, X4 | | | | | | | |
| c. 模型中的预测变量：(常量), X6, X1, X2, X3 | | | | | | | |
| d. 模型中的预测变量：(常量), X6, X1, X3 | | | | | | | |
| e. 模型中的预测变量：(常量), X1, X3 | | | | | | | |
| f. 模型中的预测变量：(常量), X1 | | | | | | | |
|  | | | | | | | |
| **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) | | 10.787 | 11.589 |  | 0.931 | 0.362 |
| X1 | | 0.613 | 0.161 | 0.671 | 3.809 | 0.001 |
| X2 | | -0.073 | 0.136 | -0.073 | -0.538 | 0.596 |
| X3 | | 0.320 | 0.169 | 0.309 | 1.901 | 0.070 |
| X4 | | 0.082 | 0.221 | 0.070 | 0.369 | 0.715 |
| X5 | | 0.038 | 0.147 | 0.031 | 0.261 | 0.796 |
| X6 | | -0.217 | 0.178 | -0.183 | -1.218 | 0.236 |
| 2 | (常量) | | 12.798 | 8.491 |  | 1.507 | 0.145 |
| X1 | | 0.613 | 0.158 | 0.671 | 3.885 | 0.001 |
| X2 | | -0.072 | 0.133 | -0.073 | -0.543 | 0.592 |
| X3 | | 0.312 | 0.162 | 0.301 | 1.924 | 0.066 |
| X4 | | 0.098 | 0.208 | 0.084 | 0.470 | 0.643 |
| X6 | | -0.211 | 0.173 | -0.178 | -1.218 | 0.235 |
| 3 | (常量) | | 14.303 | 7.740 |  | 1.848 | 0.076 |
| X1 | | 0.653 | 0.131 | 0.715 | 5.006 | 0.000 |
| X2 | | -0.077 | 0.131 | -0.077 | -0.588 | 0.562 |
| X3 | | 0.324 | 0.157 | 0.312 | 2.058 | 0.050 |
| X6 | | -0.172 | 0.149 | -0.145 | -1.151 | 0.261 |
| 4 | (常量) | | 13.578 | 7.544 |  | 1.800 | 0.084 |
| X1 | | 0.623 | 0.118 | 0.681 | 5.271 | 0.000 |
| X3 | | 0.312 | 0.154 | 0.301 | 2.026 | 0.053 |
| X6 | | -0.187 | 0.145 | -0.158 | -1.291 | 0.208 |
| 5 | (常量) | | 9.871 | 7.061 |  | 1.398 | 0.174 |
| X1 | | 0.644 | 0.118 | 0.704 | 5.432 | 0.000 |
| X3 | | 0.211 | 0.134 | 0.204 | 1.571 | 0.128 |
| 6 | (常量) | | 14.376 | 6.620 |  | 2.172 | 0.039 |
| X1 | | 0.755 | 0.098 | 0.825 | 7.737 | 0.000 |
| a. 因变量：Y | | | | | | | |

回归方程为：Y = 14.376 + 0.755\*X1

ANOVA分析知F(1,28) = 59.861, *p*<0.001, 即该回归方程模型有效。自变量对因变量的决定系数R2=0.681，调整后的R2=0.670，即Y的变异中有68.1%可由X1解释。

由表1可知X1,X2,X3,X4与Y的相关系数均显著，但只有X1能预测Y的变化，可能是因为X2,X3,X4与X1存在较高的相关。由表1知，X2,X3,X4均各自与X1为显著的高相关。再做一个X2,X3,X4关于X1的回归方程可得：

X1= 3.376 + 0.3\*X2 + 0.213\*X3 + 0.546\*X4

ANOVA分析知F(3,26) = 10.644, p<0.001, 即该回归方程模型有效。自变量对因变量的决定系数R2=0.551，调整后的R2=0.499，即X1的变异中有55.1%可由X2,X3,X4解释(X2: *t*=1.791, *p*=0.085; X3: *t*=1.047, *p*=0.305; X4: *t*=2.443, *p*=0.022)。即可知可能是因为X2,X3,X4与X1存在较高的相关，X4对X1有较高的预测性，所以它们不能预测Y变量。