第四章 回归分析

一、水泥数据

1、数据

序号 X1 X2 X3 X4 Y

1 7 26 6 60 78.5

2 1 29 15 52 74.3

3 11 56 8 20 104.3

4 11 31 8 47 87.6

5 7 52 6 33 95.9

6 11 55 9 22 109.2

7 3 71 17 6 102.9

8 1 31 22 44 72.5

9 2 54 18 22 93.1

10 21 47 4 26 115.9

11 1 40 23 34 83.8

12 11 66 9 12 113.3

13 10 68 8 12 109.4

2、经典多元线性回归分析

（1）Sas语句

data chapter4;

input number x1 x2 x3 x4 y;

cards;

1 7 26 6 60 78.5

2 1 29 15 52 74.3

3 11 56 8 20 104.3

4 11 31 8 47 87.6

5 7 52 6 33 95.9

6 11 55 9 22 109.2

7 3 71 17 6 102.9

8 1 31 22 44 72.5

9 2 54 18 22 93.1

10 21 47 4 26 115.9

11 1 40 23 34 83.8

12 11 66 9 12 113.3

13 10 68 8 12 109.4

;

proc reg;

model1 y=x1 x2 x3 x4 ;

run;

(2)sas 输出结果

REG 过程

模型: MODEL1

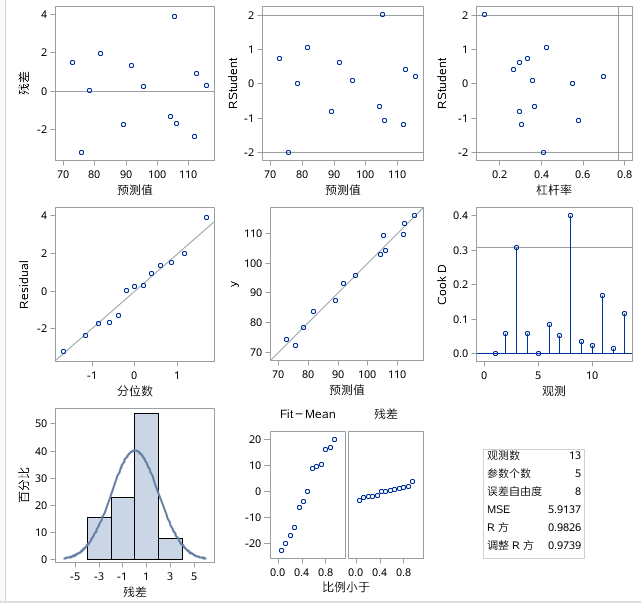
因变量: y

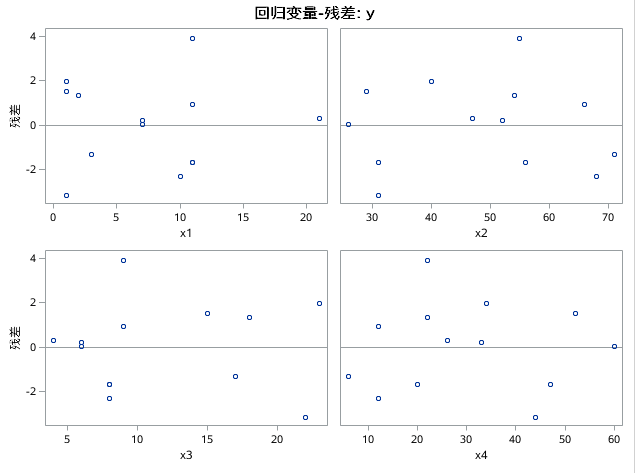
|  |  |
| --- | --- |
| **读取的观测数** | 13 |
| **使用的观测数** | 13 |

| **方差分析** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **源** | **自由度** | **平方和** | **均方** | **F 值** | **Pr > F** |
| **模型** | 4 | 2671.40128 | 667.85032 | 112.93 | <.0001 |
| **误差** | 8 | 47.30949 | 5.91369 |  |  |
| **校正合计** | 12 | 2718.71077 |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
| **均方根误差** | 2.43181 | **R 方** | 0.9826 |
| **因变量均值** | 95.43846 | **调整 R 方** | 0.9739 |
| **变异系数** | 2.54804 |  |  |

| **参数估计** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **变量** | **自由度** | **参数 估计** | **标准 误差** | **t 值** | **Pr > |t|** |
| **Intercept** | **1** | 62.06560 | 69.66415 | 0.89 | 0.3990 |
| **x1** | **1** | 1.55143 | 0.74045 | 2.10 | 0.0694 |
| **x2** | **1** | 0.51484 | 0.71959 | 0.72 | 0.4947 |
| **x3** | **1** | 0.10565 | 0.75033 | 0.14 | 0.8915 |
| **x4** | **1** | -0.14127 | 0.70494 | -0.20 | 0.8462 |





（3）sas语句说明

data chapter4; /\*建立的数据库为临时数据库work.chapter4\*/

input number x1 x2 x3 x4 y;

cards;

1 7 26 6 60 78.5

2 1 29 15 52 74.3

3 11 56 8 20 104.3

4 11 31 8 47 87.6

5 7 52 6 33 95.9

6 11 55 9 22 109.2

7 3 71 17 6 102.9

8 1 31 22 44 72.5

9 2 54 18 22 93.1

10 21 47 4 26 115.9

11 1 40 23 34 83.8

12 11 66 9 12 113.3

13 10 68 8 12 109.4

;

proc reg; /\*proc表示过程步，reg是直线回归\*/

model1 y=x1 x2 x3 x4 ;

run;

(4)sas结果分析

1、多元线性回归模型

回归方程为y=62.06560+1.55143x1+0.51484x2+0.10565x3-0.14127x4;

均方根误差为2.43181，故方差的估计量为R2=0.9826

2、回归方程和回归系数的显著性检验

平方和分解式TSS=ESS+MSS

其中TSS=2718.71077

ESS=47.30949

MSS=2671.40128

回归方程的显著性检验

F=112.93

p≤0.001

故回归方程式显著的

回归系数的显著性检验

检验统计量的值依次为0.89 2.10 0.72 0.14 -0.20、

P依次为0.3990 0.0694 0.4947 0.8519 0.8462

故 重要性依次为x1 intercept x2 x4 x3

3、逐步回归分析

（1）sas语句

data chapter4;

input number x1 x2 x3 x4 y;

cards;

1 7 26 6 60 78.5

2 1 29 15 52 74.3

3 11 56 8 20 104.3

4 11 31 8 47 87.6

5 7 52 6 33 95.9

6 11 55 9 22 109.2

7 3 71 17 6 102.9

8 1 31 22 44 72.5

9 2 54 18 22 93.1

10 21 47 4 26 115.9

11 1 40 23 34 83.8

12 11 66 9 12 113.3

13 10 68 8 12 109.4

;

proc reg;

model2 y=x1 x2 x3 x4/tol vif collin selection=stepwise r ;

run;

（2）sas输出结果

|  |
| --- |
| **逐步选择: 第 1 步** |

|  |
| --- |
| **变量 x4 已输入: R 方 = 0.6764 和 C(p) = 139.7602** |
|  |
|  |
|  |
|  |

| **方差分析** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **源** | **自由度** | **平方和** | **均方** | **F 值** | **Pr > F** |
| **模型** | 1 | 1838.98937 | 1838.98937 | 22.99 | 0.0006 |
| **误差** | 11 | 879.72140 | 79.97467 |  |  |
| **校正合计** | 12 | 2718.71077 |  |  |  |

| **变量** | **参数 估计** | **标准 误差** | **II 型 SS** | **F 值** | **Pr > F** |
| --- | --- | --- | --- | --- | --- |
| **Intercept** | 117.62615 | 5.24985 | 40148 | 502.01 | <.0001 |
| **x4** | -0.73959 | 0.15423 | 1838.98937 | 22.99 | 0.0006 |

|  |
| --- |
| **条件数字的边界: 1, 1** |

|  |
| --- |
| **逐步选择: 第 2 步** |

|  |
| --- |
| **变量 x1 已输入: R 方 = 0.9726 和 C(p) = 5.5767** |
|  |
|  |
|  |
|  |

| **方差分析** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **源** | **自由度** | **平方和** | **均方** | **F 值** | **Pr > F** |
| **模型** | 2 | 2644.33637 | 1322.16819 | 177.77 | <.0001 |
| **误差** | 10 | 74.37439 | 7.43744 |  |  |
| **校正合计** | 12 | 2718.71077 |  |  |  |

| **变量** | **参数 估计** | **标准 误差** | **II 型 SS** | **F 值** | **Pr > F** |
| --- | --- | --- | --- | --- | --- |
| **Intercept** | 103.18924 | 2.11847 | 17646 | 2372.60 | <.0001 |
| **x1** | 1.43661 | 0.13806 | 805.34701 | 108.28 | <.0001 |
| **x4** | -0.61567 | 0.04852 | 1197.59312 | 161.02 | <.0001 |

|  |
| --- |
| **条件数字的边界: 1.0641, 4.2564** |

|  |
| --- |
| **逐步选择: 第 3 步** |

|  |
| --- |
| **变量 x2 已输入: R 方 = 0.9826 和 C(p) = 3.0198** |
|  |
|  |
|  |
|  |

| **方差分析** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **源** | **自由度** | **平方和** | **均方** | **F 值** | **Pr > F** |
| **模型** | 3 | 2671.28405 | 890.42802 | 168.97 | <.0001 |
| **误差** | 9 | 47.42672 | 5.26964 |  |  |
| **校正合计** | 12 | 2718.71077 |  |  |  |

| **变量** | **参数 估计** | **标准 误差** | **II 型 SS** | **F 值** | **Pr > F** |
| --- | --- | --- | --- | --- | --- |
| **Intercept** | 71.64739 | 14.06168 | 136.80654 | 25.96 | 0.0006 |
| **x1** | 1.44863 | 0.11633 | 817.16608 | 155.07 | <.0001 |
| **x2** | 0.41734 | 0.18455 | 26.94767 | 5.11 | 0.0501 |
| **x4** | -0.23714 | 0.17230 | 9.98247 | 1.89 | 0.2020 |

|  |
| --- |
| **条件数字的边界: 18.94, 116.36** |

|  |
| --- |
| **逐步选择: 第 4 步** |

|  |
| --- |
| **变量 x4 已删除: R 方 = 0.9789 和 C(p) = 2.7079** |
|  |
|  |
|  |
|  |

| **方差分析** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **源** | **自由度** | **平方和** | **均方** | **F 值** | **Pr > F** |
| **模型** | 2 | 2661.30158 | 1330.65079 | 231.78 | <.0001 |
| **误差** | 10 | 57.40919 | 5.74092 |  |  |
| **校正合计** | 12 | 2718.71077 |  |  |  |

| **变量** | **参数 估计** | **标准 误差** | **II 型 SS** | **F 值** | **Pr > F** |
| --- | --- | --- | --- | --- | --- |
| **Intercept** | 52.52781 | 2.27638 | 3056.83522 | 532.46 | <.0001 |
| **x1** | 1.46504 | 0.12078 | 844.65627 | 147.13 | <.0001 |
| **x2** | 0.66411 | 0.04566 | 1214.55833 | 211.56 | <.0001 |

|  |
| --- |
| **条件数字的边界: 1.0551, 4.2205** |

|  |
| --- |
| **留在模型中的所有变量的显著性水平都为 0.1500。** |

|  |
| --- |
| **没有其他变量满足 0.1500 显著性水平，无法输入该模型。** |

|  |
| --- |
|  |
|  |
|  |

| **“逐步选择”的汇总** | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **步** | **输入的 变量** | **删除的 变量** | **引入 变量数** | **偏 R 方** | **模型 R 方** | **C(p)** | **F 值** | **Pr > F** |
| **1** | x4 |  | 1 | 0.6764 | 0.6764 | 139.760 | 22.99 | 0.0006 |
| **2** | x1 |  | 2 | 0.2962 | 0.9726 | 5.5767 | 108.28 | <.0001 |
| **3** | x2 |  | 3 | 0.0099 | 0.9826 | 3.0198 | 5.11 | 0.0501 |
| **4** |  | x4 | 2 | 0.0037 | 0.9789 | 2.7079 | 1.89 | 0.2020 |

| **方差分析** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **源** | **自由度** | **平方和** | **均方** | **F 值** | **Pr > F** |
| **模型** | 2 | 2661.30158 | 1330.65079 | 231.78 | <.0001 |
| **误差** | 10 | 57.40919 | 5.74092 |  |  |
| **校正合计** | 12 | 2718.71077 |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
| **均方根误差** | 2.39602 | **R 方** | 0.9789 |
| **因变量均值** | 95.43846 | **调整 R 方** | 0.9747 |
| **变异系数** | 2.51054 |  |  |

| **参数估计** | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **变量** | **自由度** | **参数 估计** | **标准 误差** | **t 值** | **Pr > |t|** | **容差** | **方差 膨胀** |
| **Intercept** | **1** | 52.52781 | 2.27638 | 23.08 | <.0001 | . | 0 |
| **x1** | **1** | 1.46504 | 0.12078 | 12.13 | <.0001 | 0.94775 | 1.05513 |
| **x2** | **1** | 0.66411 | 0.04566 | 14.55 | <.0001 | 0.94775 | 1.05513 |

| **共线性诊断** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **个数** | **特征值** | **条件 指数** | **偏差比例** | | |
| **Intercept** | **x1** | **x2** |
| **1** | 2.70512 | 1.00000 | 0.01092 | 0.03902 | 0.01072 |
| **2** | 0.24996 | 3.28974 | 0.05486 | 0.95976 | 0.04826 |
| **3** | 0.04493 | 7.75977 | 0.93422 | 0.00121 | 0.94102 |

| **输出统计量** | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **观测** | **因 变量** | **预测 值** | **标准 误差 均值 预测** | **残差** | **残差 标准误差** | **Student 残差** | **Cook D** |
| **1** | 78.5 | 80.0498 | 1.2009 | -1.5498 | 2.073 | -0.747 | 0.062 |
| **2** | 74.3 | 73.2519 | 1.2262 | 1.0481 | 2.059 | 0.509 | 0.031 |
| **3** | 104.3 | 105.8331 | 0.8262 | -1.5331 | 2.249 | -0.682 | 0.021 |
| **4** | 87.6 | 89.2305 | 1.1793 | -1.6305 | 2.086 | -0.782 | 0.065 |
| **5** | 95.9 | 97.3165 | 0.6928 | -1.4165 | 2.294 | -0.618 | 0.012 |
| **6** | 109.2 | 105.1690 | 0.8130 | 4.0310 | 2.254 | 1.788 | 0.139 |
| **7** | 102.9 | 104.0744 | 1.4412 | -1.1744 | 1.914 | -0.614 | 0.071 |
| **8** | 72.5 | 74.5801 | 1.1767 | -2.0801 | 2.087 | -0.997 | 0.105 |
| **9** | 93.1 | 91.3196 | 1.0141 | 1.7804 | 2.171 | 0.820 | 0.049 |
| **10** | 115.9 | 114.5065 | 1.7770 | 1.3935 | 1.607 | 0.867 | 0.306 |
| **11** | 83.8 | 80.5571 | 1.0278 | 3.2429 | 2.164 | 1.498 | 0.169 |
| **12** | 113.3 | 112.4742 | 1.0625 | 0.8258 | 2.148 | 0.385 | 0.012 |
| **13** | 109.4 | 112.3373 | 1.1089 | -2.9373 | 2.124 | -1.383 | 0.174 |

（3）结果分析

1、变量的选择问题

我们选的是逐步回归的方法，由sas的输出结果可以看到，在选自变量的过程中，第一步引入变量x4，第二步引入变量x2，第三步入变量x1，第四步删除变量x4。最后回归方程还有两个自变量x1和x2。

此时，回归方程为y=52.52781+1.46504x1+0.66401x2。

2、关于变量选择的几个准则