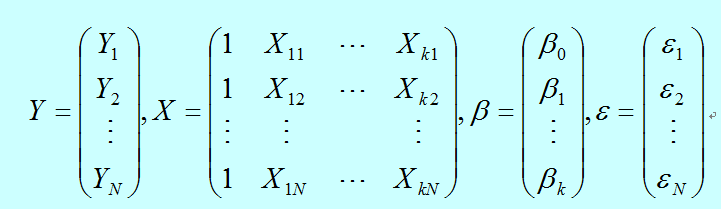
**25. 多元线性回归**

**一、多元线性回归**

**1. 模型为**

Y=𝛽0+𝛽1X1+…+ 𝛽NXN+ε

其中 X1, …, XN是自变量，Y是因变量，𝛽0, 𝛽1…, 𝛽N是待求的未知参数，ε是随机误差项（残差），若记

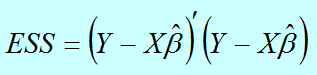


多元线性回归模型可写为矩阵形式：

Y=Xβ+ε

通常要求：矩阵X的秩为k+1（保证不出现共线性）, 且k<N; ε为正态分布，E(ε)=0 和E(εε’)= 𝜎2I，其中I为N×N单位矩阵。

用最小二乘法原理，令残差平方和



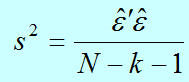
最小，得到



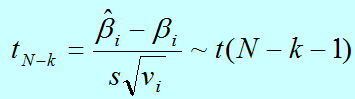
为β的最佳线性无偏估计量（高斯－马尔可夫定理）。

**2.** 𝜎2**的估计和T检验**

选取𝜎2的估计量：



则



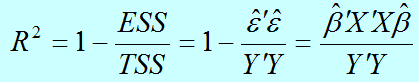
假如t值的绝对值相当大，就可以在适当选定的置信水平上否定原假设，参数的1-α置信区间可由下式得出：



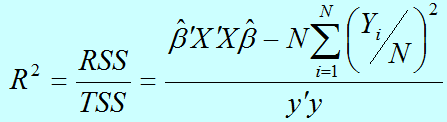
其中tα/2为与α%显著水平有关的t分布临界值。

**3. R2和F检验**





若因变量不具有0平均值，则必须对R2做如下改进：



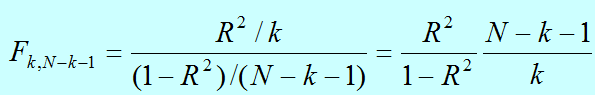
随着模型中增添新的变量，R2的值必定会增大，为了去掉这种增大的干扰，还需要对R2进行修正（校正拟合优度对自由度的依赖关系）：



做假设检验：

H0: 𝛽1=…=𝛽N=0; H1: 𝛽1…, 𝛽N至少有一个≠0；

使用F统计量做检验，



若F值较大，则否定原假设。

**二、PROC REG过程步**

基本语法：

PROC REG data = 数据集;

MODEL *因变量 = 自变量列表 </可选项>*;

< restrict 自变量的等式约束;>

说明：MODEL语句用来指定因变量和自变量；

restrict语句示例：restrict a1+a2=1;

常用的输出可选项：

STB——输出标准化偏回归系数矩阵

CORRB——输出参数估计矩阵

COLLINOINT——对自变量进行共线性分析

P——输出个体观测值、预测值及残差 （R/CLM/CLI包含P）

R——输出每个个体观测值、残差及标准误差

CLM——输出因变量均值95%的置信界限的上下限

CLI——对各预测值输出95%的置信界限的上下限

MSE——要求输出随机扰动项方差𝜎2的估计

与残差分析有关的可选项

 VIF——输出变量间相关性的方差膨胀系数，VIF越大，说明由于共线性存在，使方差变大；

COLLIN——输出条件数，它表示最大的特征值与每个自变量特征值之比的平方根。一般情况下，条件数越大越可能存在共线性；

TOL——表示共线性水平的容许值，TOL越小说明其可用别的自变量解释的部分多，自然可能与别的自变量存在共线性关系；

DW——输出Durbin-Watson统计量；

influence——对异常点进行诊断，对每一观测点输出统计量（Cook’s D > 50%, defits/debetas > 2说明该点影响较大）。

交互式语句

 add——向模型中增加变量；

 delete——删除原拟合模型中的有关变量；

 refit——重新拟合模型；

 print——输出有关模型的相关信息。

绘制回归分析的图形

在PROC REG过程步加入绘图选项语句即可。

基本语法：

PROC REG data = 数据集PLOTS *=* (*图形类型*);

可选的绘图类型：

FITPLOT——带回归线、置信预测带的散点图；

RESIDUALS——自变量的残差图；

DIAGNOSTICS——诊断图（包括下面各图）；

COOKSD——Cook's D统计量图；

OBSERVEDBYPREDICTED——根据预测值的因变量图；

QQPLOT ——检验残差正态性的QQ图；

RESIDUALBYPREDICTED——根据预测值的残差图；

RESIDUALHISTOGRAM——残差的直方图；

RFPLOT——残差拟合图；

RSTUDENTBYLEVERAGE——杠杆比率的学生化残差图；

RSTUDENTBYPREDICTED——预测值的学生化残差图；

注：残差图(RESIDUALS)和诊断图(DIAGNOSTICS)是自动生成的，根据模型也有其它默认的图形输出；若只绘制指定的图形需要加上ONLY:

**PROC** **REG** data = 数据集 PLOTS(ONLY) = (图形类型);

**例1** 用多元线性回归模型，来研究耗氧量的是如何依赖其它变量的。

31位成年人心肺功能的调查数据（见下表），由于回归是相关的，所以理论上还应该做共线性诊断。

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| age | weight | Oxygen  耗氧量 | runtime  跑15英哩的时间（分） | rstpulse  休息时每分钟心跳次数 | runpulse  跑步时每分钟心跳次数 | Maxpulse  每分钟心跳次数最大值 |
| 44 | 89.47 | 44.609 | 11.37 | 62 | 178 | 182 |
| 40 | 75.07 | 45.313 | 10.07 | 62 | 185 | 185 |
| 44 | 85.84 | 54.297 | 8.65 | 45 | 156 | 168 |
| 42 | 68.15 | 59.571 | 8.17 | 40 | 166 | 172 |
| 38 | 89.02 | 49.874 | 9.22 | 55 | 178 | 180 |
| 47 | 77.45 | 44.811 | 11.63 | 58 | 176 | 176 |
| 40 | 75.98 | 45.681 | 11.95 | 70 | 176 | 180 |
| 43 | 81.19 | 49.091 | 10.85 | 64 | 162 | 170 |
| 44 | 81.42 | 39.442 | 13.08 | 63 | 174 | 176 |
| 38 | 81.87 | 60.055 | 8.63 | 48 | 170 | 186 |
| 44 | 73.03 | 50.541 | 10.13 | 45 | 168 | 168 |
| 45 | 87.66 | 37.388 | 14.03 | 56 | 186 | 192 |
| 45 | 66.45 | 44.754 | 11.12 | 51 | 176 | 176 |
| 47 | 79.15 | 47.273 | 10.60 | 47 | 162 | 164 |
| 54 | 83.12 | 51.855 | 10.33 | 50 | 166 | 170 |
| 49 | 81.42 | 49.156 | 8.95 | 44 | 180 | 185 |
| 51 | 69.63 | 40.836 | 10.95 | 57 | 168 | 172 |
| 51 | 77.91 | 46.672 | 10.00 | 48 | 162 | 168 |
| 48 | 91.63 | 46.774 | 10.25 | 48 | 162 | 164 |
| 49 | 73.37 | 50.388 | 10.08 | 76 | 168 | 168 |
| 57 | 73.37 | 39.407 | 12.63 | 58 | 174 | 176 |
| 54 | 79.38 | 46.080 | 11.17 | 62 | 156 | 165 |
| 52 | 76.32 | 45.441 | 9.63 | 48 | 164 | 166 |
| 50 | 70.87 | 54.625 | 8.92 | 48 | 146 | 155 |
| 51 | 67.25 | 45.118 | 11.08 | 48 | 172 | 172 |
| 54 | 91.63 | 39.203 | 12.88 | 44 | 168 | 172 |
| 51 | 73.71 | 45.790 | 10.47 | 59 | 186 | 188 |
| 57 | 59.08 | 50.545 | 9.93 | 49 | 148 | 155 |
| 49 | 76.32 | 48.673 | 9.40 | 56 | 186 | 188 |
| 48 | 61.24 | 47.920 | 11.50 | 52 | 170 | 176 |
| 52 | 82.78 | 47.467 | 10.50 | 53 | 170 | 172 |

**代码：**

**data** fitness ;

input age weight oxygen runtime rstpulse runpulse maxpulse;

datalines;

44 89.47 44.609 11.37 62 178 182

40 75.07 45.313 10.07 62 185 185

44 85.84 54.297 8.65 45 156 168

42 68.15 59.571 8.17 40 166 172

38 89.02 49.874 9.22 55 178 180

47 77.45 44.811 11.63 58 176 176

40 75.98 45.681 11.95 70 176 180

43 81.19 49.091 10.85 64 162 170

44 81.42 39.442 13.08 63 174 176

38 81.87 60.055 8.63 48 170 186

44 73.03 50.541 10.13 45 168 168

45 87.66 37.388 14.03 56 186 192

45 66.45 44.754 11.12 51 176 176

47 79.15 47.273 10.60 47 162 164

54 83.12 51.855 10.33 50 166 170

49 81.42 49.156 8.95 44 180 185

51 69.63 40.836 10.95 57 168 172

51 77.91 46.672 10.00 48 162 168

48 91.63 46.774 10.25 48 162 164

49 73.37 50.388 10.08 76 168 168

57 73.37 39.407 12.63 58 174 176

54 79.38 46.080 11.17 62 156 165

52 76.32 45.441 9.63 48 164 166

50 70.87 54.625 8.92 48 146 155

51 67.25 45.118 11.08 48 172 172

54 91.63 39.203 12.88 44 168 172

51 73.71 45.790 10.47 59 186 188

57 59.08 50.545 9.93 49 148 155

49 76.32 48.673 9.40 56 186 188

48 61.24 47.920 11.50 52 170 176

52 82.78 47.467 10.50 53 170 172

;

**run**;

**proc** **corr** data = fitness PLOT = MATRIX(HISTOGRAM nvar=all);

var oxygen age weight runtime rstpulse runpulse maxpulse;

label oxygen = 'Oxygen consumption'

age = 'Age in years'

weight = 'weight in kg'

runtime = 'Min. to run 1.5 miles'

rstpulse = 'Heart rate while resting'

runpulse = 'Heart rate while running'

maxpulse = 'Maximum heart rate';

**run**;

**proc** **reg** data = fitness PLOTS(ONLY) = (DIAGNOSTICS FITPLOT);

model oxygen = age maxpulse rstpulse runpulse runtime weight/ss1 ss2; /\* ss1为第Ⅰ类型平方和， ss2为第Ⅱ类型平方和 \*/

**run**;

delete rstpulse;

print;

**run**;

**proc** **reg** data= fitness;

model oxygen = age maxpulse runpulse runtime weight;

pulse: test maxpulse+runpulse=**0**;

**run**;

**proc** **reg** data= fitness;

model oxygen = age maxpulse runpulse runtime weight/ss2; /\* 带restrict约束条件的回归，ss1不可用 \*/

restrict maxpulse+runpulse=**0**;

**run**;

**data** fitness2;

set fitness;

maxrun=maxpulse-runpulse;

**run**;

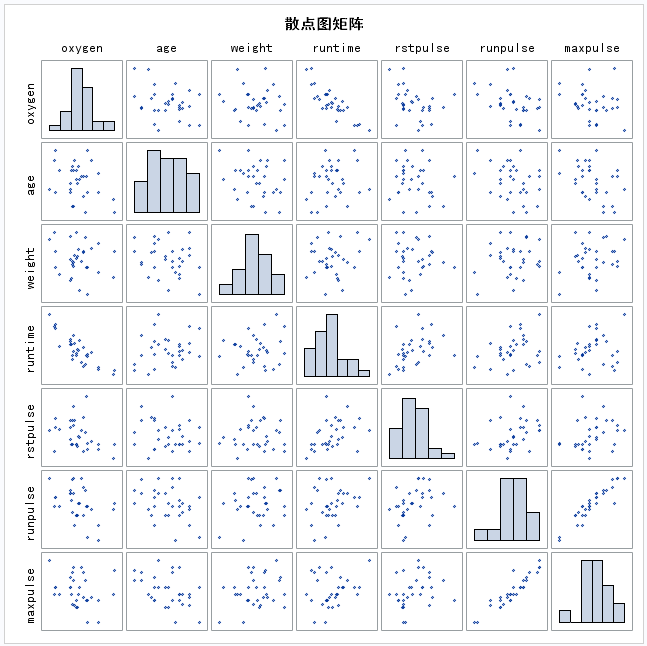
**proc** **reg** data= fitness2;

model oxygen = age maxrun runtime weight/ss1 ss2;

**run**;

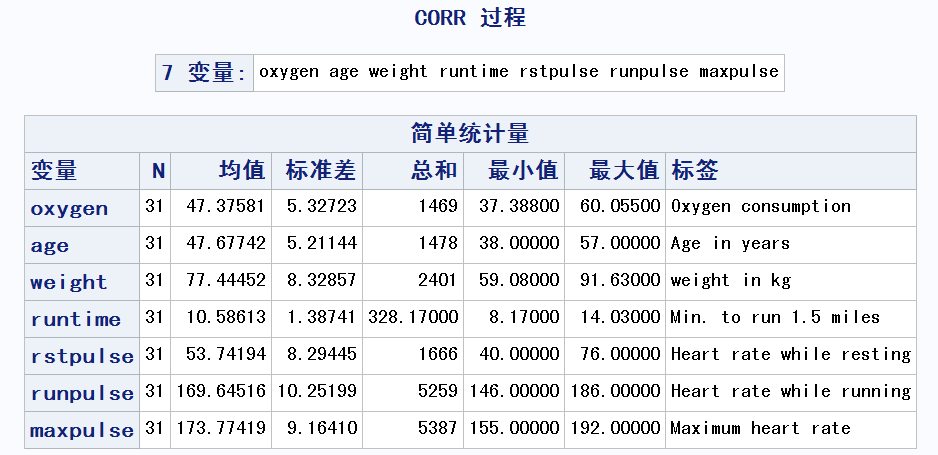
运行结果及说明：

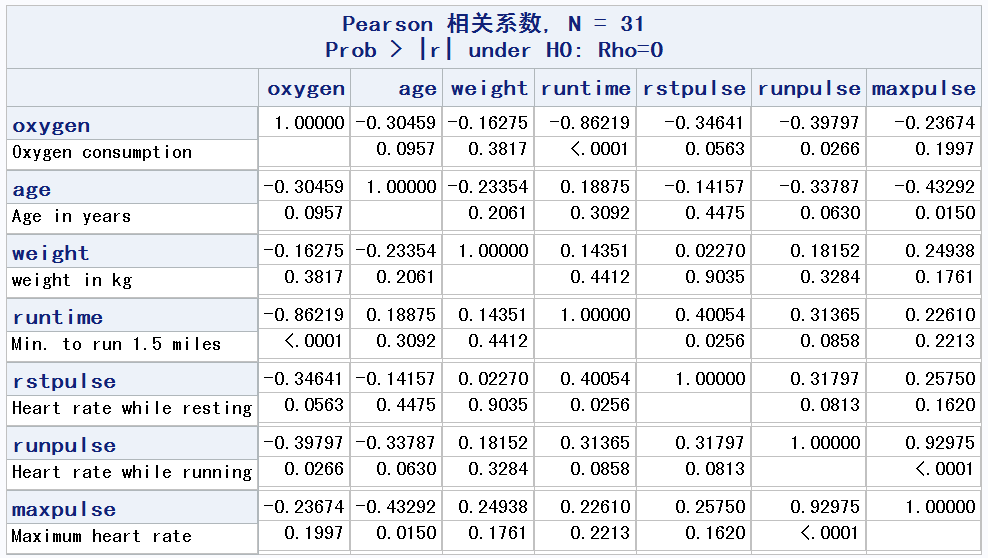
（1）相关性分析



散点图矩阵中第一行的6个散点图分别表示oxygen变量作为y轴,其他六个变量作为x轴的散点图,第一列的6个散点图分别表示oxygen变量作为x轴,其他六个变量作为y轴的散点图；对角线是该变量数值变化的直方图。

绘制散点图矩阵图是为了观察变量间的相关性。从图中可以看出变量runpulse与maxpulse之间存在有较强的共线性，如果在回归模型中增加方差膨胀系数（vif），共线性水平的容许值（tol），条件数（collin）选项对回归进行共线性诊断，也会得到相同的结论。另外，我们从图中还发现耗氧量oxygen与变量runtime有较强的负相关，从下面的相关系数也能得到相同的结论。





（2）回归分析





得到回归方程：

oxygen = 102.238339-0.219916age + 0.304735maxpulse-0.000844rstpuls

-0.373164runpulse-2.680516runtime-0.072380weight

多元线性回归模型的一个重要问题是，如何正确地缩减自变量到达最优的简化模型。

判断回归模型是否还能缩减自变量，可以通过这第Ⅰ类平方和（I型SS）和第Ⅱ类平方和（Ⅱ型SS），构造F检验（等同于这个自变量的参数t检验，因为F=t2）来比较确定。

runtime自变量的两类平方和都是最大的且占的比例很大，说明是回归模型中第一重要的自变量。而rstpulse自变量在第一类平方和中有比较大的数值却在第Ⅱ类平方和中是最小的，这是rstpulse自变量（休息时每分钟心跳次数）应该被考虑第一个删除的主要原因。

（3）剔除不显著的回归变量，继续回归分析

自变量rstpulse和weight的回归系数的t检验的P值分别为0.9886和0.1980，都大于α=0.05，故不拒绝“系数=0”的原假设。

这里要小心地看待这些检验，因为它们都是在其他自变量都加入回归的前提下进行显著性检验的，完全可能因为自变量间存在较强的相关而掩盖他们对回归的贡献。

所以，在剔除不显著的回归变量时必须逐个进行。另外，从自变量rstpulse的回归系数更接近于0，也提示我们应先考虑删除自变量rstpulse.

delete rstpulse;

print;

**run**;





得到新的回归方程：

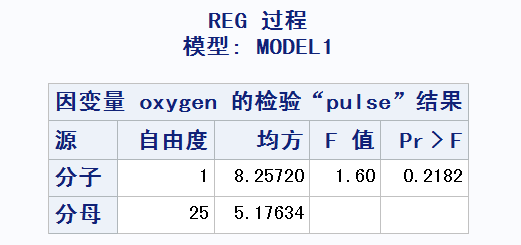
oxygen = 102.204275-0.219621age+0.304908maxpulse

-0.373401runpulse-2.682523runtime-0.072302weight

比较剔除rstpulse前后的输出可看出，R2的变化在4位小数之后。进入回归的变量的回归系数在剔除rstpulse后的变化都小于 1％。说明剔除自变量rstpulse后，对回归模型没有实质性的影响，这个自变量的剔除是恰当的。

（4）系数的线性组合检验

使用test语句，还可以对回归系数的线性组合进行F检验。 例如，我们看到参数估计表中，maxpulse变量的系数与runpulse变量的系数之和非常接近于0（**回顾前文：这两个自变量有很强的共线性**），想要检验一下这样的系数线性组合是否有很大的可能性，如果是那样，则可以简化回归模型。



在test语句中的线性方程组只有1个，所以自由度为1，线性方程组的平方和为8.2572。F检验的分子为均方和8.2572/1=8.2572，F检验的分母为均方误差MSE=5.176338，自由度为25。因此，

F(1,25)=8.2572/5.176338=1.5952，P值= 0.2182>0.05，

即不能拒绝回归系数线性组合的原假设。

（5）有约束条件的回归模型

若回归模型中的回归系数有线性组合的约束条件，则可以使用restrict语句在此约束条件下重新拟合回归模型。

**proc** **reg** data= fitness;

model oxygen=age maxpulse runpulse runtime weight/ss2; /\* 带restrict约束条件的回归，ss1不可用 \*/

restrict maxpulse+runpulse=**0**;

**run**;





得到新的回归方程：

oxygen = 89.487420-0.162973age + 0.388260(maxpulse-runpulse)

-2.805552 runtime-0.080635weight

注意到maxpulse变量的系数为0.388260, runpulse变量的系数为－0.388260, 两者之和为0. 同时，它们的第Ⅱ类平方和也相等为57.444946.

没有约束条件时的RSS=721.97309，增加约束条件后的RSS= 713.71590，减少了8.25720，说明为了增加这个约束条件损失了回归平方和8.25720，但换来了简化的回归模型。这个数值越小越表明这个约束条件客观存在。所以restrict语句的第Ⅱ类平方和为8.257198，且自由度为－1（是减少不是增加）。同时注意到这个8.25720数字就是前面线性组合检验结果中的分子值。

自变量系数存在线性组合的约束条件，本质上是指这些自变量可以通过这种组合方式转换为一个新的自变量，例如，在这个样本中产生一个新变量maxrun=maxpulse-runpulse，这个新变量maxrun加入到回归模型中，而变量maxpulse和runpulse从回归模型中删除，重新拟合回归模型。

**data** fitness2;

set fitness;

maxrun=maxpulse-runpulse;

**run**;

**proc** **reg** data= fitness2;

model oxygen = age maxrun runtime weight/ss1 ss2;

**run**;

运行结果与前面相同。





回归方程为：

oxygen = 89.487420-0.162973age + 0.388260maxrun

-2.805552 runtime-0.080635weight