Regression analysis

吴文韬
10165000225
统计

一、Multiple regression and prediction

I choose The FITNESS data for this analysis. It is from a study of how fast the body can absorb and use up oxygen. Data were collected as 31 subjects, in three groups, were passed through an exercise run.

These measurements were made on men involved in a physical fitness course at N.C.State Univ. The variables and their lables are shown as table following: (See appendix I for detailed data)

column	type	label	column	type	label
AGE	num	Age in years	RUNPULSE	num	Heart rate while running
GROUP	num	Experimental group	RSTPULSE	num	Heart rate while resting
MAXPULSE	num	Maximum heart rate	RUNTIME	num	Min. to run 1.5 miles
OXYGEN	num	Oxygen consumption	WEIGHT	num	Weight in kg

For this data set, we want to model the Oxygen consumption based on the varibles Age, Weight, RunTime, RunPulse, RestPulse and MaxPulse. The general model would be $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$, with $\boldsymbol{\epsilon} \sim N\left(0, \sigma^2 \boldsymbol{I_n}\right)$, $\boldsymbol{Y}$ is $(\boldsymbol{Y_1}, \boldsymbol{\ldots}, \boldsymbol{Y_n})'$, $\boldsymbol{X}$ is $(\boldsymbol{x_1}, \boldsymbol{\ldots}, \boldsymbol{x_n})'$, and $\boldsymbol{\beta}$ means $(\beta_0, \beta_1, ..., \beta_p)'$. Using this data set, we have 31 observations, $p = 6$ and $\boldsymbol{x_i} = (1, x_{i1}, \cdots, x_{i6})'$. So we creat the model:

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \beta_5 x_{5i} + \beta_6 x_{6i} + \epsilon_i$$

where $(i = 1, \cdots, 31)$, and $(x_1, \cdots, x_6)$ represent $Age, Weight, RunTime, RunPulse, RestPulse, MaxPulse$.

Then we use SAS to get the analysis results:

REG 过程
模型: MODEL1
因变量: Oxygen

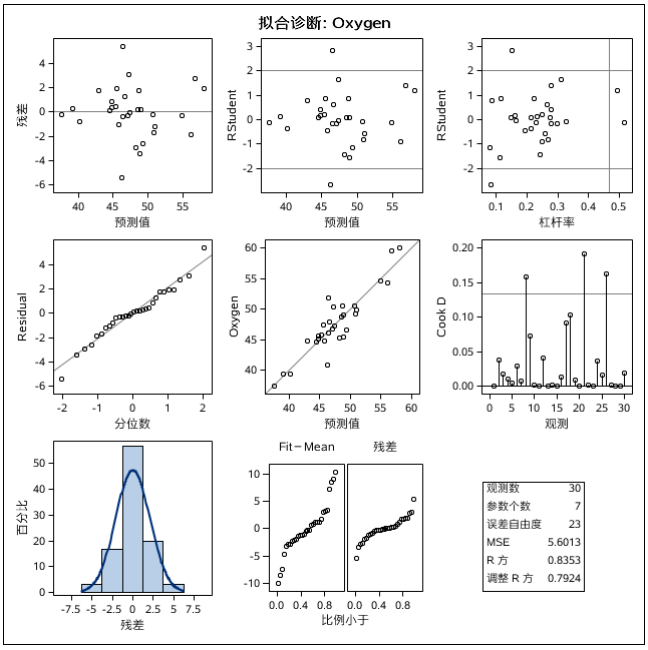
读取的观测数	31
使用的观测数	31

方差分析					
源	自由度	平方和	均方	F 值	Pr > F
模型	6	722.54361	120.42393	22.43	<.0001
误差	24	128.83794	5.36825		
校正合计	30	851.38154			

均方根误差	2.31695	R 方	0.8487
因变量均值	47.37581	调整 R 方	0.8108
变异系数	4.89057		

参数估计					
变量	自由度	参数估计	标准误差	t 值	Pr > |t|
Intercept	1	102.93448	12.40326	8.30	<.0001
Age	1	-0.22697	0.09984	-2.27	0.0322
Weight	1	-0.07418	0.05459	-1.36	0.1869
RunTime	1	-2.62865	0.38456	-6.84	<.0001
RunPulse	1	-0.36963	0.11985	-3.08	0.0051
RestPulse	1	-0.02153	0.06605	-0.33	0.7473
MaxPulse	1	0.30322	0.13650	2.22	0.0360

(a) Result of reg



(b) Predicted and residuals

图 1 Result of model

From the figure above we can find the estimate of β and get the model: (Keep three valid

Numbers)

$$Y = 102.934 - 0.227\,x_1 - 0.074\,x_2 - 2.629\,x_3 - 0.37\,x_4 - 0.021\,x_5 + 0.303\,x_6$$

though the P-value of variable RestPulse($Pr = 0.7473$) and Weight($Pr = 0.1869$) are a little big. From the qqplot of residuals(Residuals versus quantile) we can find that the assumption of normal probability of error is reasonable. And the prediction is shown as this table:

Obs	Oxygen	Predicted Value	Std Error Mean Predict	95%CL Mean		95%CL Predict		Residual	Std Error Residual	Student Residual	Cook's D
1	44.6	44.4799	0.8734	42.6773	46.2825	39.3695	49.5903	0.1291	2.146	0.060	0.000
2	54.3	56.1519	1.1443	53.7903	58.5136	50.8186	61.4853	-1.8549	2.015	-0.921	0.039
3	49.9	51.0710	1.1743	48.6474	53.4946	45.7100	56.4321	-1.1970	1.997	-0.599	0.018
4	45.7	44.8244	1.1907	42.3668	47.2819	39.4479	50.2009	0.8566	1.988	0.431	0.010
5	39.4	40.2197	1.0438	38.0654	42.3740	34.9749	45.4645	-0.7777	2.069	-0.376	0.005
6	50.5	48.7762	1.0925	46.5213	51.0311	43.4893	54.0631	1.7648	2.043	0.864	0.030
7	44.8	45.7745	1.0219	43.6654	47.8835	40.5481	51.0008	-1.0205	2.079	-0.491	0.008
8	51.9	46.4703	0.8320	44.7531	48.1875	41.3894	51.5512	5.3847	2.162	2.490	0.131
9	40.8	46.2386	0.6639	44.8684	47.6087	41.2642	51.2129	-5.4026	2.220	-2.434	0.076
10	46.8	47.1135	1.1692	44.7003	49.5266	41.7571	52.4698	-0.3395	2.000	-0.170	0.001
11	39.4	39.1567	1.0656	36.9574	41.3560	33.8933	44.4202	0.2503	2.057	0.122	0.001
12	45.4	48.8382	0.7598	47.2701	50.4063	43.8057	53.8707	-3.3972	2.189	-1.552	0.041
13	45.1	44.7887	0.9153	42.8996	46.6777	39.6471	49.9302	0.3293	2.128	0.155	0.001
14	45.8	45.3528	1.1495	42.9804	47.7252	40.0147	50.6909	0.4372	2.012	0.217	0.002
15	48.7	48.4904	1.1986	46.0167	50.9641	43.1065	53.8743	0.1826	1.983	0.092	0.000
16	47.5	45.5659	0.7433	44.0318	47.1001	40.5439	50.5879	1.9011	2.194	0.866	0.012
17	45.3	48.1954	1.0672	45.9928	50.3981	42.9306	53.4603	-2.8824	2.057	-1.402	0.076
18	59.6	56.8041	1.2075	54.3119	59.2963	51.4117	62.1965	2.7669	1.977	1.399	0.104
19	44.8	43.0132	0.6705	41.6294	44.3970	38.0351	47.9914	1.7978	2.218	0.811	0.009
20	49.1	48.9203	1.0635	46.7255	51.1152	43.6587	54.1819	0.1707	2.058	0.083	0.000
21	60.1	58.0793	1.6245	54.7266	61.4321	52.2392	63.9195	1.9757	1.652	1.196	0.198
22	37.4	37.5993	1.5015	34.5004	40.6982	31.9010	43.2976	-0.2113	1.765	-0.120	0.001
23	47.3	47.3677	0.8839	45.5435	49.1919	42.2496	52.4858	-0.0947	2.142	-0.044	0.000
24	49.2	50.8615	1.1855	48.4147	53.3083	45.4899	56.2331	-1.7055	1.991	-0.857	0.037
25	46.7	49.3203	0.6103	48.0607	50.5798	44.3752	54.2653	-2.6483	2.235	-1.185	0.015
26	50.4	47.2738	1.2236	44.7485	49.7991	41.8660	52.6816	3.1142	1.968	1.583	0.138
27	46.1	46.4614	1.2187	43.9463	48.9766	41.0584	51.8645	-0.3814	1.971	-0.194	0.002
28	54.6	54.8806	1.1055	52.5990	57.1623	49.5822	60.1790	-0.2556	2.036	-0.126	0.001
29	39.2	39.1324	1.4192	36.2033	42.0615	33.5246	44.7401	0.0706	1.831	0.039	0.000
30	50.5	50.7506	1.3230	48.0201	53.4812	45.2440	56.2573	-0.2056	1.902	-0.108	0.001
31	47.9	46.6774	1.1923	44.2166	49.1381	41.2994	52.0553	1.2426	1.987	0.626	0.020

二、Testing linear hypotheses

We can notice that the P-value of variable RestPulse and Weight are kind of inappropriate, so the hypothesis we test is $H_0 : \beta_{Weight} = \beta_{RestPulse} = 0$, the result is:

The SAS System

The REG Procedure
Model: MODEL1

Test 1 Results for Dependent Variable Oxygen				
Source	DF	Mean Square	F Value	Pr > F
Numerator	2	5.04604	0.94	0.4045
Denominator	24	5.36825		

图 2　Result of hypothesis test

From the table we can find that the P-value is 0.4045 > 0.05, which means we can not reject the hypothesis, $i.e.$, the effect of Weight and RestPulse can be ignored.

三、Estimation under linear restrictions and generalized least squares

In the former section we talk about testing the hypothsis H_0, and the conclusion is we do not reject it, which means the contribution of RestPulse and Weight to the dependent variable is insignificant, so we add this restriction: $\boldsymbol{A\beta = c}$, where the $\boldsymbol{A}$ is $(\boldsymbol{a_1}, \boldsymbol{a_2})'$, $\boldsymbol{a_1} = (0,1,0,0,0,0)'$, $\boldsymbol{a_2} = (0,0,0,0,1,0)'$, $\boldsymbol{c} = (0,0)'$. and the result is:

SAS 系统

REG 过程
模型: MODEL1
因变量: Oxygen

Note: Restrictions have been applied to parameter estimates.

读取的观测数	31
使用的观测数	31

方差分析					
源	自由度	平方和	均方	F 值	Pr > F
模型	4	712.45153	178.11288	33.33	<.0001
误差	26	138.93002	5.34346		
校正合计	30	851.38154			

均方根误差	2.31159	R 方	0.8368
因变量均值	47.37581	调整 R 方	0.8117
变异系数	4.87927		

(a) Result of reg

参数估计					
变量	自由度	参数估计	标准误差	t 值	Pr > |t|
Intercept	1	98.14789	11.78569	8.33	<.0001
Age	1	-0.19773	0.09564	-2.07	0.0488
Weight	1	1.36741E-17	0	正无	<.0001
RunTime	1	-2.76758	0.34054	-8.13	<.0001
RunPulse	1	-0.34811	0.11750	-2.96	0.0064
RestPulse	1	-1.7347E-18	0	负无穷	<.0001
MaxPulse	1	0.27051	0.13362	2.02	0.0533
RESTRICT	-1	-131.69100	98.65367	-1.33	0.1871*
RESTRICT	-1	-15.02684	81.53621	-0.18	0.8580*

* 使用 beta 分布计算的概率。

(b) Estimation of parameter

图 3　Result of model with restriction

From the table we can get the new equation:

$$Y = 98.148 - 0.198\,x_1 - 2.768\,x_3 - 0.348\,x_4 + 0.271\,x_6$$

which means the consumption of oxygen is positively correlated with maxpulse, and negatively correlated with age, runtime and runpulse.

Now we consider this model: $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \boldsymbol{\epsilon} \sim \left(0, \sigma^2 V\right)$, where V is a positive semi-definite matrix. If $V > 0$, the model can be rewritten as

$$V^{-\frac{1}{2}}\boldsymbol{Y} = V^{-\frac{1}{2}}\boldsymbol{X}\boldsymbol{\beta} + V^{-\frac{1}{2}}\boldsymbol{\epsilon}, \boldsymbol{\epsilon} \sim \left(0, \sigma^2 V\right)$$

or

$$\tilde{\boldsymbol{Y}} = \tilde{\boldsymbol{X}}\boldsymbol{\beta} + \tilde{\boldsymbol{\epsilon}}, \tilde{\boldsymbol{\epsilon}} \sim \left(0, \sigma^2 \boldsymbol{I_n}\right)$$

so we use a WEIGHT statement in SAS. Assume that the data are simulated to have variance proportional to $Oxygen^2$, the result of analysis with this weighted data set is:

SAS 系统

REG 过程
模型: MODEL1
因变量: Oxygen

读取的观测数	31
使用的观测数	31

权重: Weights

方差分析

源	自由度	平方和	均方	F 值	Pr > F
模型	6	0.31215	0.05203	21.78	<.0001
误差	24	0.05733	0.00239		
校正合计	30	0.36949			

均方根误差	0.04888	R 方	0.8448
因变量均值	46.24956	调整 R 方	0.8060
变异系数	0.10568		

(a) Result of reg

参数估计

变量	自由度	参数估计	标准误差	t 值	Pr > |t|
Intercept	1	101.73436	12.37232	8.22	<.0001
Age	1	-0.22126	0.09859	-2.24	0.0343
Weight	1	-0.06176	0.05430	-1.14	0.2666
RunTime	1	-2.59861	0.35583	-7.30	<.0001
RunPulse	1	-0.35546	0.12724	-2.79	0.0101
RestPulse	1	-0.01820	0.06530	-0.28	0.7829
MaxPulse	1	0.28531	0.14336	1.99	0.0581

(b) Estimation of parameter

图 4 Generalized least squares

四、Test ANOVA

In this section we use a new data set:

The quality control department of a fabric finishing plant is studying the effect of several factors on the dyeing of cotton-synthetic cloth used to manufacture men's shirts. Three operators, three cycle times, and two temperatures were selected, and three small specimens of cloth were dyed under each set of conditions. The finished cloth was compared to a standard, and a numerical score was assigned. The results are as follows.

Cycle Time	Temperature					
	300°C			350°C		
	Operator			Operator		
	1	2	3	1	2	3
40	23	27	31	24	38	34
	24	28	32	23	36	36
	25	26	29	28	35	39
50	36	34	33	37	34	34
	35	38	34	39	38	36
	36	39	35	35	36	31
60	28	35	26	26	36	28
	24	35	27	29	37	26
	27	34	25	25	34	24

The score is the dependent variable, and Temperature, Operator and Cycle time are independent variables with 2 (300°C and 350°C), 3 (Operator1, 2 and 3) and 3 (40, 50 and 60) levels respectively. Then we get the anova table:

SAS 系统

ANOVA 过程

因变量: Score Score

源	自由度	平方和	均方	F 值	Pr > F
模型	17	1239.333333	72.901961	22.24	<.0001
误差	36	118.000000	3.277778		
校正合计	53	1357.333333			

R 方	变异系数	均方根误差	Score 均值
0.913065	5.737384	1.810463	31.55556

源	自由度	Anova 平方和	均方	F 值	Pr > F
Cycle_Time	2	436.0000000	218.0000000	66.51	<.0001
Temperature	1	50.0740741	50.0740741	15.28	0.0004
Cycle_Tim*Temperatur	2	78.8148148	39.4074074	12.02	0.0001
Operator	2	261.3333333	130.6666667	39.86	<.0001
Cycle_Time*Operator	4	355.6666667	88.9166667	27.13	<.0001
Temperature*Operator	2	11.2592593	5.6296296	1.72	0.1939
Cycle_*Temper*Operat	4	46.1851852	11.5462963	3.52	0.0159

图 5 Anova Table

From the table we can find that execpt $Temperature * Operator$, other source of effect satisfy the significant level. That means the main effect of $Temperature$, $Operator$ and $Cycletime$, and the interaction from $Cycletime * Operator$, $Temperature * Operator$ and $Cycletime * Temperature * Operator$ are significant. So if we want to create a model, it would be (estimates of parameter are omitted)

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_{12} x_1 x_2 + \beta_{13} x_1 x_3 + \beta_{123} x_1 x_2 x_3$$

APPENDIX I: FITNESS data

Age	Weight	Oxygen	RunTime	RestPulse	RunPulse	MaxPulse
44	89.47	44.609	11.37	62	178	182
44	85.84	54.297	8.65	45	156	168
38	89.02	49.874	9.22	55	178	180
40	75.98	45.681	11.95	70	176	180
44	81.42	39.442	13.08	63	174	176
44	73.03	50.541	10.13	45	168	168
45	66.45	44.754	11.12	51	176	176
54	83.12	51.855	10.33	50	166	170
51	69.63	40.836	10.95	57	168	172
48	91.63	46.774	10.25	48	162	164
57	73.37	39.407	12.63	58	174	176
52	76.32	45.441	9.63	48	164	166
51	67.25	45.118	11.08	48	172	172
51	73.71	45.790	10.47	59	186	188
49	76.32	48.673	9.40	56	186	188
52	82.78	47.467	10.50	53	170	172
40	75.07	45.313	10.07	62	185	185
42	68.15	59.571	8.17	40	166	172
47	77.45	44.811	11.63	58	176	176
43	81.19	49.091	10.85	64	162	170
38	81.87	60.055	8.63	48	170	186
45	87.66	37.388	14.03	56	186	192
47	79.15	47.273	10.60	47	162	164
49	81.42	49.156	8.95	44	180	185
51	77.91	46.672	10.00	48	162	168
49	73.37	50.388	10.08	67	168	168
54	79.38	46.080	11.17	62	156	165
50	70.87	54.625	8.92	48	146	155
54	91.63	39.203	12.88	44	168	172
57	59.08	50.545	9.93	49	148	155
48	61.24	47.920	11.50	52	170	176

APPENDIX II: SAS code

```
libname mylib "C:\Users\35150\Documents\Regression";
/* import the data */
proc import out = mylib.regression
Datafile = "C:\Users\35150\Desktop\data.txt" dbms = dlm
replace;
Getnames = YES;
run;quit;
/* view the data */
proc print data = mylib.regression; run;quit;
proc contents varnum data=mylib.regression;
   ods select position;
run;
/* ods word */
ods rtf file="C:\Users\35150\Documents\My SAS Files\reg.doc";
ods graphics on;
/* linear regression */
proc reg data = mylib.regression alpha = 0.05
         /* make the plot of residual with smooth line */
         plots = (Residuals(smooth) DFBETAS ObservedByPredicted(label));
model Oxygen=Age Weight RunTime RunPulse RestPulse MaxPulse/ r cli clm ;
/* r cli clm is to show the predicted value and residual*/
/* restrict Age + Weight + RunTime + RunPulse + RestPulse + MaxPulse = -3;*/
/* test Weight = RestPulse = 0;*/
output out = result_out
predicted = Fitted
LCLM=Lower Est UCLM=Upper Est LCL=Lower Pred UCL=Upper Pred
residual = Residuals;
ods output ParameterEstimates = mylib_parameters;
ods output ANOVA = mylib_anova;
run;quit;
ods graphics off; ods rtf close;
/*proc anova data = mylib.regression ;
class Age Weight RunTime RunPulse RestPulse MaxPulse;
model Oxygen=Age Weight RunTime RunPulse RestPulse MaxPulse;
run;quit;*/
/* test hypothesis*/
proc reg data = mylib.regression alpha = 0.05;
model Oxygen=Age Weight RunTime RunPulse RestPulse MaxPulse;
/*mtest Weight , RestPulse;*/
test Weight =0 , RestPulse = 0;
run;quit;
/* add restriction */
proc reg data = mylib.regression alpha = 0.05;
```

```
model Oxygen=Age Weight RunTime RunPulse RestPulse MaxPulse;
/*delete Weight RestPulse;*/
restrict Weight =0, RestPulse = 0;
run;quit;
/* generalized least squares */
data a;
set mylib.regression;
Weights = 1/Oxygen**2;
run;
/* view data a */
proc print data = a;run;
proc reg data = a;
model Oxygen=Age Weight RunTime RunPulse RestPulse MaxPulse;
weight Weights;
run;quit;
/* import the anova data*/
proc import out = mylib.anova
Datafile = "C:\Users\35150\Desktop\data2.xls" dbms = excel
replace;
Getnames = YES;
run;quit;
/* view the data */
proc print data = mylib.anova; run;quit;
/* anova */
proc anova data = mylib.anova;
class Cycle_Time Temperature Operator;
model Score = Cycle_Time | Temperature | Operator;
run;quit;
```