The George Washington University
6210,2018 spring

Report of Homework5

Jiaying Liu

March 28, 2018

1 Data set

The data set contain one response variable and six explanatory variables. We want to explore the relation response variable between Oxygen intake rate and the explanatory variables age,weight, runtime, restpulse, runpulse and maxpulue. The runtime represents the time to run 1.5 miles. The restpluse represents the heart rate while resting. The runpulse represents the heart rate while running. The maxpulse represents the maximum heart rate recorded while running. There are 31 individual observation. We are interested in exploring the significant impact of each explanatory variables and we also want to find a parsimonious model to best explain the oxygen intake rate base on those significant variables.

Table 1: Original Data set.

Obs	**Age**	**Weight**	**Oxygen**	**RunTime**	**RestPulse**	**RunPulse**	**MaxPulse**
1	44	89.47	44.609	11.37	62	178	182
2	40	75.07	45.313	10.07	62	185	185
3	44	85.84	54.297	8.65	45	156	168
4	42	68.15	59.571	8.17	40	166	172
5	38	89.02	49.874	9.22	55	178	180
6	47	77.45	44.811	11.63	58	176	176
7	40	75.98	45.681	11.95	70	176	180
8	43	81.19	49.091	10.85	64	162	170
9	44	81.42	39.442	13.08	63	174	176
...	...	...	...	...	...	...	...
28	57	59.08	50.545	9.93	49	148	155
29	49	76.32	48.673	9.40	56	186	188
30	48	61.24	47.920	11.50	52	170	176
31	52	82.78	47.467	10.50	53	170	172

2 Significant explanatory variables

2.1 Perform a simple linear regression

In the begin we just fit the simple linear regression which only contain one explanatory variable. We can see that the p value of variable age, weight and maxpulse is greater than 0.05 which mean those variable are not significant if we fit the one variable linear regression. But since the oxygen intake rate will definitely depend on several variable. So we need the examine it in more complex situation.

Table 2: Simple linear regression between each variables and response

Model	**F Value**	**Pr>F**	**variable**	**t value**	**Pr>t**	**R-Square**	**Adj R-Sq**
Oxygen = Age	2.97	0.0957	Age	-1.72	0.0957	0.0928	0.0615
Oxygen = Weight	0.79	0.3817	Weight	-0.89	0.3817	0.0265	0.007
Oxygen = Runtime	84.01	<.0001	RunTime	-9.17	<.0001	0.7435	0.7345
Oxygen = RestPulse	5.50	0.0260	RestPulse	-2.35	0.0260	0.1595	0.1305
Oxygen = RunPulse	5.46	0.0266	RunPulse	-2.34	0.0266	0.1584	0.1294
Oxygen = MaxPulse	1.72	0.1997	MaxPulse	-1.31	0.1997	0.0560	0.0235

2.2 Type III SS

From the table 3 which is the result of type III SS table. As we know the type III SS is one of the critical testing to verify the significant of variables. From the type III SS we can explore whether the variable is important or not when it is adding to the model which already contains all other variable.From table 3, we can find the p value of weight variable and the restpulse variable is greater than 0.05 which mean we can not reject the original hypothesis. The weight and restpulse is not significantly impact on oxygen intake rate.

Table 3: Type III SS.

Source	**DF**	**type III SS**	**Mean Square**	**Fvalue**	**Pr>F**
Age	1	27.7457715	27.7457715	5.17	0.0322
Weight	1	9.9105884	9.9105884	1.85	0.1869
RunTime	1	250.8221009	250.8221009	46.72	<.0001
RestPulse	1	0.5705130	0.5705130	0.11	0.7473
RunPulse	1	51.0580583	51.0580583	9.51	0.0051
MaxPulse	1	26.4914241	26.4914241	4.93	0.0360

2.3 conclusion

From the above analysis, we can raise the answer of first problem. We can conclude that the age, runtime, runpulse and maxpulse are the significant variable. They have obviously impact on the oxygen intake rate.

3 Correlation

As we know, the effect of highly correlated explanatory variable are confounded.If there are strong correlation among several explanatory variable, we need to delete it to import the preform of model. So in the following steo we use the pearson correlation coefficient and spearmen correlation coefficients to check the relation of each individual explanatory variable and use the VIF to check the multicollinearity.

First, let's consider the correlation between each explanatory variable. From the table 4, it is the result of peasrson correlation coefficients. In order to avoid the error of non-normal distribution, we also perform the result of spearman correlation coefficients(table 5). Table's result show the p value of index. If we found the p value is smaller than 0.05, we have reason to believe such two variable have some linear relation. Combine this two table, we can easily conclude that age and maxpulse have linear relation, runtime and runpulse have linear relation, runpulse and maxpulse have strong linear relation.

Table 4: Pearson Correlation Coefficients,.

source	**Age**	**Weight**	**RunTime**	**RestPulse**	**RunPulse**	**MaxPulse**
Age	1	0.2061	0.3092	0.3777	0.063	0.015
Weight	0.2061	1	0.4412	0.8143	0.3284	0.1761
RunTime	0.3092	0.4412	1	0.011	0.0858	0.2213
RestPulse	0.3777	0.8143	0.011	1	0.0518	0.0951
RunPulse	0.063	0.3284	0.0858	0.0518	1	<.0001
MaxPulse	0.015	0.1761	0.2213	0.0951	<.0001	1

Table 5: Spearman Correlation Coefficients,.

source	**Age**	**Weight**	**RunTime**	**RestPulse**	**RunPulse**	**MaxPulse**
Age	1	0.3853	0.3934	0.5285	0.1033	0.0316
Weight	0.3853	1	0.6891	0.8745	0.6868	0.4441
RunTime	0.3934	0.6891	1	0.0056	0.1217	0.2667
RestPulse	0.5285	0.8745	0.0056	1	0.0419	0.0732
RunPulse	0.1033	0.6868	0.1217	0.0419	1	<.0001
MaxPulse	0.0316	0.4441	0.2667	0.0732	<.0001	1

3.1 VIF multicollinearity

Multicollinearity is a statistical phenomenon in which two or more explanatory variables in a mutiple regression model are highly correlated. Usually we use indicator VIF to check the multicollinearity. VIF of an explanatory variable indicates the strength of the linear relationship between the variable and the remaining ones.Normally, if the VIF greater than 10, we will consider that the variable strongly related with other variable. From the table 6, it show the result of VIF. We can find that none of variable's VIF greater than 10 which mean there are no exact one variable are muti-relate with the other variable.

Table 6: VIF.

variable	label	df	parameter estimate	SE	t value	Pr>t	variable inflation
Intercept	Intercept	1	102.93448	12.40326	8.30	<.0001	0
Age	Age	1	-0.22697	0.09984	-2.27	0.0322	1.51284
Weight	Weight	1	-0.07418	0.05459	-1.36	0.1869	1.15533
RunTime	RunTime	1	-2.62865	0.38456	-6.84	<.0001	1.59087
RestPulse	RestPulse	1	-0.02153	0.06605	-0.33	0.7473	1.41559
RunPulse	RunPulse	1	-0.36963	0.11985	-3.08	0.0051	8.43727
MaxPulse	MaxPulse	1	0.30322	0.13650	2.22	0.0360	8.74385

4 Parsimonious model

4.1 All possible subset regression

From above analysis, we already analyze the significant of variable and the correlation of variable. We find that there are four variable are significant and there are not multicollinear variable. So in the following step, we begin to find the parsimonious model. There are several way to find the model. Firstly, we list all possible subset of parsimonious model and use the indicators Mallow's statistic c(p) and the R-Squares for the models examined.Then we will find the model with C(p) is equal to p. And then when c(p) = P, then we will choose the smallest p. From the figure 1, it show the all possible model's c(p) and p, we are going to find the point which c(p) = p. From the figure 1 we find that the line cross the c(p)∗p point on p equal to 5 and 7. And since we need to find the smallest c(p), so when c(p) = 5 is the best model which contains age weight runtime runpulse and maxpulse 5 variables.

Figure 1: Scatter Plot of Cp vs. P

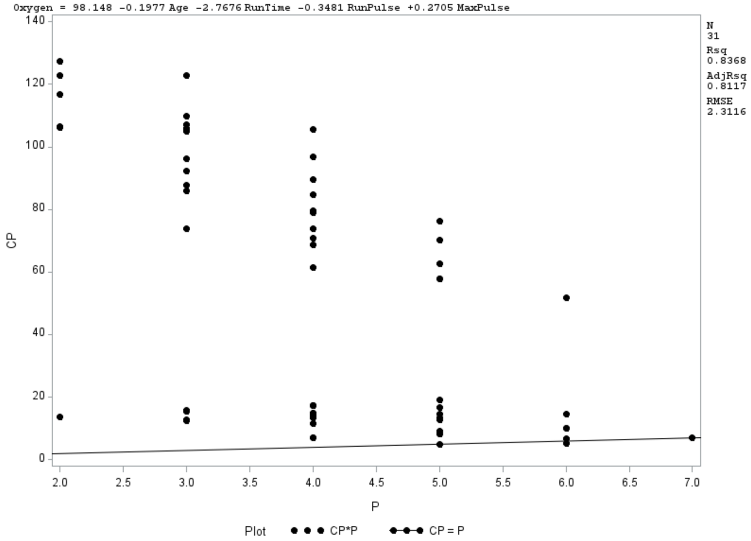


Table 7: C(p) = p.

Number in Model	**C(P)**	**R-Squares**	**Variables in Model**
5	5.1063	0.8480	Age Weight RunTime RunPulse MaxPulse
6	7.0000	0.8487	Age Weight RunTime RestPulse RunPulse MaxPulse

4.2 Stepwise selection method

Let's consider another selection method. The forward selection. In this approach, one adds variables to the model one at a time. At each step, each variable that is not already in the model is tested for inclusion in the model. The most significant of these variables is added to the model, so long as it's P-value is below some pre-set level. Thus we begin with a model including the variable that is most significant in the initial analysis, and continue adding variables until none of remaining variables are "significant" when added to the model. Note that this multiple use of hypothesis testing means that the real type I SS for a variable. The table 8 show the summary of forward selection. In the first step, it include the most significant variable runtime. And the adding Age, Runpulse, maxpulse and the weight. However, as we mention above, the forward select is similar like type I SS table. It really depend on the order of variable adding. So we still need another selection to find the best parsimonious model.

Table 8: Summary of Forward Selection.

Step	**Entered**	**vars number**	**Partial R^2**	**Model R^2**	**C(p)**	**F value**	**Pr>F**
1	RunTime	1	0.7434	0.7434	13.6988	84.01	<.0001
2	Age	2	0.0209	0.7642	12.3894	2.48	0.1267
3	RunPulse	3	0.0468	0.8111	6.9596	6.70	0.0154
4	MaxPulse	4	0.0257	0.8368	4.8800	4.10	0.0533
5	Weight	5	0.0112	0.8480	5.1063	1.84	0.1871

Forward selection has drawbacks, including the fact that each addition of a new variable may render one or more of the already included variables non-significant. An alternate approach which avoids this is backward selection. Under this approach, one starts with fitting a model with all the variables of interest (following the initial screen). Then the least significant variable is dropped, so long as it is not significant at our chosen critical level. We continue by successively re-fitting reduced models and applying the same rule until all remaining variables are statistically significant.So from the table 10, we can find that the backward selection removed the restpulse and weight variable which are no significant variable.

Table 9: Summary of Backward Selection.

Step	**Removed**	**var number**	**Partial R^2**	**Model R^2**	**C(p)**	**F value**	**Pr>F**
1	RestPulse	5	0.0007	0.8480	5.1063	0.11	0.7473
2	Weight	4	0.0112	0.8368	4.8800	1.84	0.1871

Stepwise selection is a method that allows moves in either direction, dropping or adding variables at the various steps. Backward stepwise selection involves starting off in a backward approach and then potentially adding back variables if they later appear to be significant. The process is one of alternation between choosing the least significant variable to drop and then re-considering all dropped variables (except the most recently dropped) for re-introduction into the model. This means that two separate significance levels must be chosen for deletion from the model and for adding to the model. The second significance must be more stringent than the first.So from the table 11, displaying the result if stepwise select. It is show that the variable runtime age runpulse and maxpulse are added and there are no remove variable.

Table 10: Summary of Stepwise Selection.

Step	**Entered**	**Removed**	**var number**	**Partial R^2**	**Model R^2**	**C(p)**	**F value**	**Pr>F**
1	RunTime		1	0.7434	0.7434	13.6988	84.01	<.0001
2	Age		2	0.0209	0.7642	12.3894	2.48	0.1267
3	RunPulse		3	0.0468	0.8111	6.9596	6.70	0.0154
4	MaxPulse		4	0.0257	0.8368	4.8800	4.10	0.0533

4.3 Conclusion

From the above analysis of significant variable and the different method of selection. we can conclude that the best parsimonious model is only contain four variable which are age, runtime,runpulse and the maxpulse.

$$Oxygen = \\ 98.14789 - 0.19773Age - 2.73758Runtime - 0.34811RunPulse + 0.27051MaxPulse$$

The model is significant with the p value smaller than .0001, and each variable are significant with p value smaller than 0.05.

Table 11: Analysis of Variance.

Source	**DF**	**Sum of Square**	**Mean Square**	**F value**	**P > f**
Model	4	712.45153	178.11288	33.33	<.0001
Error	26	138.93002	5.34346		
Corrected Total	30	851.38154			

Table 12: Parameter Estimates.

Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t value**	**P > t**
Intercept	1	98.14789	11.78569	8.33	<.0001
Age	1	-0.19773	0.09564	-2.07	0.0488
RunTime	1	-2.76758	0.34054	-8.13	<.0001
RunPulse	1	-0.34811	0.11750	-2.96	0.0064
MaxPulse	1	0.27051	0.13362	2.02	0.0533

5 Verify the assumption

since we have found the best parsimonious model, the final and most important thing is to verify several assumption.

First,we need to make sure the error term is independent in each explanatory variables and response variable(y hat). From the figure 2, it show the the scatter plot of residuals with each explanatory variables. It is obviously that the mean of residual is zero and there are no pattern between residual and variables. The residual is a horizontal line and we can find the dots of explanatory variable are randomly disperse below or above the line of residual. So we can conclude that the error is independent in explanatory variables.

Secondly, let's verify the error with response variables. From the figure 3,it is the result of error with predict y. Same as above, the residual is a horizontal pattern and the predict y is randomly disperse above or below the residual. So again, we can conclude that the residual is independent in fitting values.

Thirdly, we need to verify that the residual is normal distribution. From the figure 5, displaying the QQ plot of residual. The linear line is so clear that we have reason to believe the residual have normal distribution.

In the end, we need to check whether there contain outliers. From the figure 6 which is the result of cook's distance. If the cook's distance is large, we have the reason to believe such observation are outliers. We do find a point which cook's distance is larger than any other. So if we want the more exactly result, the observation 10 must be deleted.

Figure 2: Residual Plot of each variable

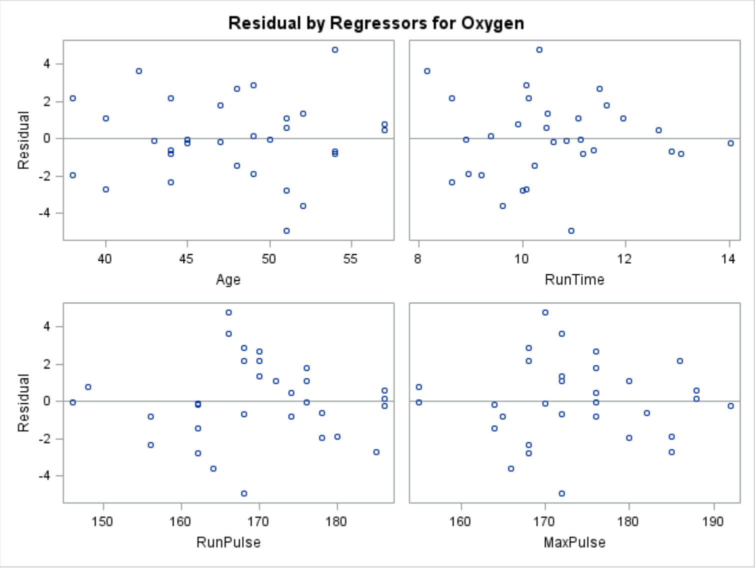


Figure 3: Residual Plot predict

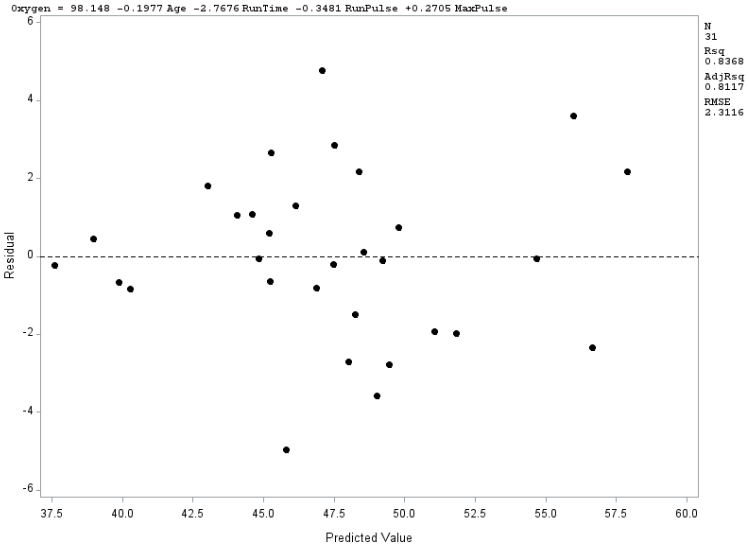


Figure 4: normality of residual(PP PLOT)

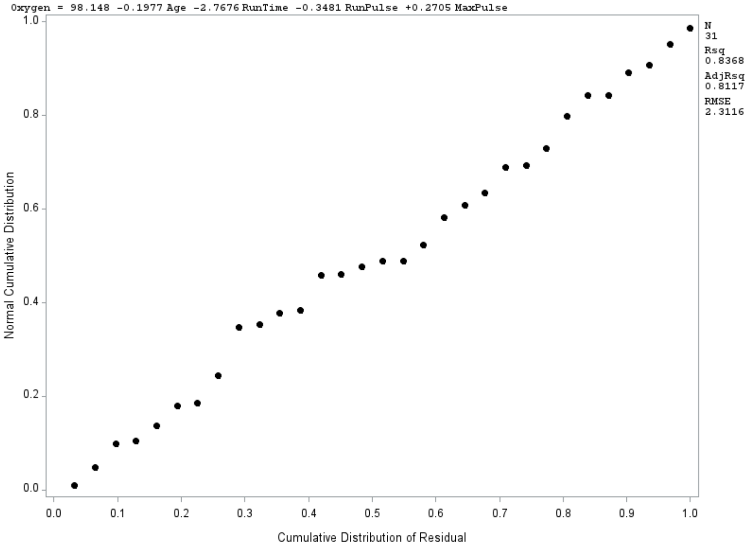


Figure 5: normality of residual(QQ PLOT)

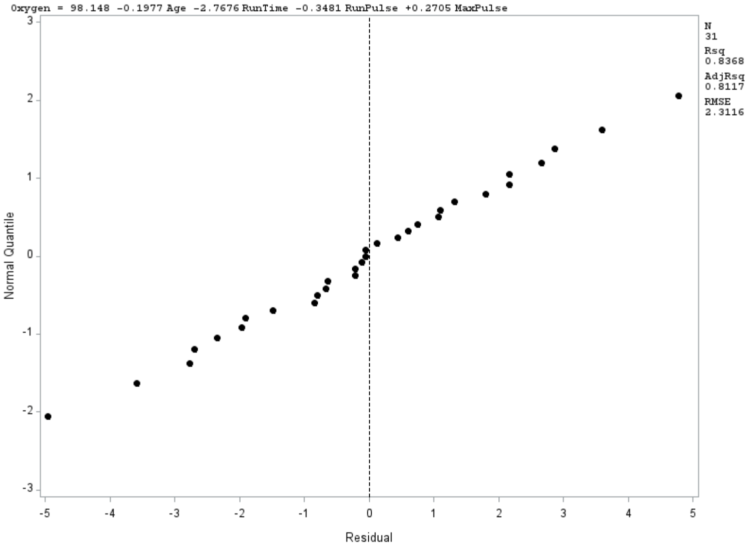
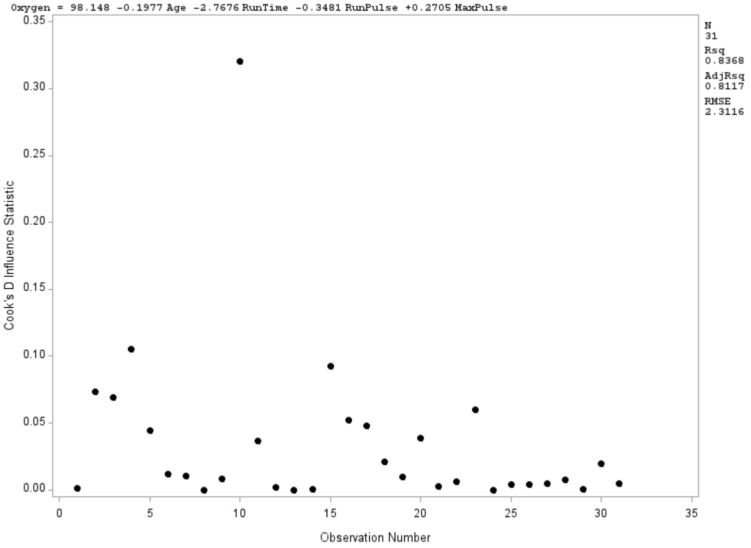


Figure 6: Cook's distance


Oxygen = 98.148 -0.1977 Age -2.7676 RunTime -0.3481 RunPulse +0.2705 MaxPulse
N
Rsq
AdjRsq
RMSE
Cook's D Influence Statistic
Observation Number