HW 5 Report

Yongchao Qiao

I Perform simple linear regressions

1. Model 1: Oxygen ~ Age

Table 1: Analysis of variance

Source	DF	Sum of squares	Mean Square	F Value	Pr > F
Model	1	78.98823	78.98823	2.97	0.0957
Error	29	772.39332	26.63425		
Corrected Total	30	851.38154			

Table 2: Parameter Estimates

Variable	DF	Parameter Estimates	Standard Error	t Value	Pr > |t|
Intercept	1	62.22064	8.66983	7.18	<0.0001
Age	1	-0.31136	0.18080	-1.72	0.0957

Table 3: Statistical indicators

Root MSE	Dependent mean	Coeff Var	R-Square	Adj R-sq
5.16084	47.37581	10.89340	0.0928	0.0615

The table 1 shows the p-value of the F test is greater than 0.05, which indicates the simple linear regression model is not significant to describe the relationship between variable Oxygen and variable Age. Besides, the table 2 shows the p-value of the t test for the estimated coefficient of Age is greater than 0.05, which indicates the variable Age is not significantly to explain the change of variable Oxygen. Then the table 3 shows the R-Square as 0.0928, which is quite small, indicates variable Age can hardly explain the change of variable Oxygen.

2. Model 2: Oxygen ~ Weight

Table 4: Analysis of variance

Source	DF	Sum of squares	Mean Square	F Value	Pr > F
Model	1	22.55181	22.55181	0.79	0.3817
Error	29	828.82973	28.58034		
Corrected Total	30	851.38154			

Table 5: Parameter Estimates

Variable	DF	Parameter Estimates	Standard Error	t Value	Pr > |t|
Intercept	1	55.43795	9.12663	6.07	<0.0001
Weight	1	-0.10410	0.11719	-0.89	0.3817

Table 6: Statistical indicators

Root MSE	Dependent mean	Coeff Var	R-Square	Adj R-sq
5.34606	47.37581	11.28436	0.0265	-0.0071

The table 4 shows the p-value of the F test is greater than 0.05, which indicates the simple linear regression model is not significant to describe the relationship between variable Oxygen and variable Weight. Besides, the table 5 shows the p-value of the t test for the estimated coefficient of Weight is greater than 0.05, which indicates the variable Weight is not significantly to explain the change of variable Oxygen. Then the table 6 shows the R-Square as 0.0265, which is quite small, indicates variable Weight can hardly explain the change of variable Oxygen.

3. Model 3: Oxygen ~ RunTime

Table 7: Analysis of variance

Source	DF	Sum of squares	Mean Square	F Value	Pr > F
Model	1	632.90010	632.90010	84.01	<0.0001
Error	29	218.48144	7.53384		
Corrected Total	30	851.38154			

Table 8: Parameter Estimates

Variable	DF	Parameter Estimates	Standard Error	t Value	Pr > |t|
Intercept	1	82.42177	3.85530	21.38	<0.0001
RunTime	1	-3.31056	0.36119	-9.17	<0.0001

Table 9: Statistical indicators

Root MSE	Dependent mean	Coeff Var	R-Square	Adj R-sq
2.74478	47.37581	5.79364	0.7434	0.7345

The table 7 shows the p-value of the F test is less than 0.05, which indicates the simple linear regression model is significant to describe the relationship between variable Oxygen and variable RunTime. Besides, the table 8 shows the p-value of the t test for the estimated coefficient of RunTime is less than 0.05, which indicates the variable RunTime is significantly to explain the change of variable Oxygen. Then the table 9 shows the R-Square as 0.7434,

which is not that small, indicates variable RunTime can explain the change of variable Oxygen in a certain degree.

4. Model 4: Oxygen ~ RestPulse

Table 10: Analysis of variance

Source	DF	Sum of squares	Mean Square	F Value	Pr > F
Model	1	135.78285	135.78285	5.50	0.0260
Error	29	715.59870	24.67582		
Corrected Total	30	851.38154			

Table 11: Parameter Estimates

Variable	DF	Parameter Estimates	Standard Error	t Value	Pr > |t|
Intercept	1	62.30029	6.42453	9.70	<0.0001
RestPulse	1	-0.27921	0.11903	-2.35	0.0260

Table 12: Statistical indicators

Root MSE	Dependent mean	Coeff Var	R-Square	Adj R-sq
4.96748	47.37581	10.48526	0.1595	0.1305

The table 10 shows the p-value of the F test is less than 0.05, which indicates the simple linear regression model is significant to describe the relationship between variable Oxygen and variable RestPulse. Besides, the table 11 shows the p-value of the t test for the estimated coefficient of RestPulse is less than 0.05, which indicates the variable RestPulse is significantly to explain the change of variable Oxygen. Then the table 12 shows the R-Square as 0.1595, which is not that big, indicates variable RestPulse can explain the change of variable Oxygen in a certain degree.

5. Model 5: Oxygen ~ RunPulse

Table 13: Analysis of variance

Source	DF	Sum of squares	Mean Square	F Value	Pr > F
Model	1	134.84474	134.84474	5.46	0.0266
Error	29	716.53681	24.70817		
Corrected Total	30	851.38154			

Table 14: Parameter Estimates

Variable	DF	Parameter Estimates	Standard Error	t Value	Pr > |t|
Intercept	1	82.45825	15.04386	5.48	<0.0001
RunPulse	1	-0.20680	0.08852	-2.34	0.0266

Table 15: Statistical indicators

Root MSE	Dependent mean	Coeff Var	R-Square	Adj R-sq
4.97073	47.37581	10.49213	0.1584	0.1294

The table 13 shows the p-value of the F test is less than 0.05, which indicates the simple linear regression model is significant to describe the relationship between variable Oxygen and variable RunPulse. Besides, the table 14 shows the p-value of the t test for the estimated coefficient of RunPulse is less than 0.05, which indicates the variable RunPulse is significantly to explain the change of variable Oxygen. Then the table 15 shows the R-Square as 0.1584, which is not that big, indicates variable RunPulse can explain the change of variable Oxygen in a certain degree.

6. Model 6: Oxygen ~ MaxPulse

Table 16: Analysis of variance

Source	DF	Sum of squares	Mean Square	F Value	Pr > F
Model	1	47.71646	47.71646	1.72	0.1997
Error	29	803.66508	27.71259		
Corrected Total	30	851.38154			

Table 17: Parameter Estimates

Variable	DF	Parameter Estimates	Standard Error	t Value	Pr > |t|
Intercept	1	71.29074	18.24976	3.91	0.0005
MaxPulse	1	-0.13762	0.10488	-1.31	0.1997

Table 18: Statistical indicators

Root MSE	Dependent mean	Coeff Var	R-Square	Adj R-sq
5.26427	47.37581	11.11174	0.0560	0.0235

The table 16 shows the p-value of the F test is greater than 0.05, which indicates the simple linear regression model is not significant to describe the relationship between variable Oxygen and variable MaxPulse. Besides, the table 17 shows the p-value of the t test for the estimated coefficient of MaxPulse is greater than 0.05, which indicates the variable MaxPulse is not significantly to explain the change of variable Oxygen. Then the table 18 shows the R-Square as 0.0560, which is quite small, indicates variable MaxPulse can hardly explain the change of variable Oxygen.

Therefore, in these simple linear regressions, only variable RunTime, RunPulse and RestPulse have significant impacts on the oxygen intake rate.

II Perform a correlation analysis for all the explanatory variables

1. Pearson's correlation coefficient

$H_0 : Rho = 0, \quad H_1 : Rho \neq 0$

Table 19: Pearson Correlation Test and Pearson's correlation coefficient

	Age	Weight	RunTime	RestPulse	RunPulse	MaxPulse
Age	1.00000	-0.23354 (0.2061)	0.18875 (0.3092)	-0.16410 (0.3777)	-0.33787 (0.0630)	-0.43292 (**0.0150**)
Weight	-0.23354 (0.2061)	1.00000	0.14351 (0.4412)	0.04397 (0.8143)	0.18152 (0.3284)	0.24938 (0.1761)
RunTime	0.18875 (0.3092)	0.14351 (0.4412)	1.00000	0.45038 (**0.0110**)	0.31365 (0.0858)	0.22610 (0.2213)
RestPulse	-0.16410 (0.3777)	0.04397 (0.8143)	0.45038 **(0.0110)**	1.00000	0.35246 (0.0518)	0.30512 (0.0951)
RunPulse	-0.33787 (0.0630)	0.18152 (0.3284)	0.31365 (0.0858)	0.35246 (0.0518)	1.00000	0.92975 (**<.0001**)
MaxPulse	-0.43292 **(0.0150)**	0.24938 (0.1761)	0.22610 (0.2213)	0.30512 (0.0951)	0.92975 **(<.0001)**	1.00000

The result is shown as table 19. The Pearson correlation coefficient between Age and MaxPulse is -0.43292 with P value less than 0.05. Thus, it indicates that the null hypothesis that Rho = 0 can be rejected. That is, variable Age and variable MaxPulse have a significant negative correlation relationship, with Pearson correlation coefficient as -0.43292. The Pearson correlation coefficient between RunTime and RestPulse is 0.45038 with P value less than 0.05. Thus, it indicates that the null hypothesis that Rho = 0 can be rejected. That is, variable RunTime and variable RestPulse have a significant positive correlation relationship, with Pearson correlation coefficient as 0.45038. The Pearson correlation coefficient between RunPulse and MaxPulse is 0.92975 with P value less than 0.05. Thus, it indicates that the null hypothesis that Rho = 0 can be rejected. That is, variable RunPulse and variable MaxPulse have a highly significant positive correlation relationship, with Pearson correlation coefficient as 0.92975.

1. Spearman's correlation coefficient

$H_0 : Rho = 0, \quad H_1 : Rho \neq 0$

Table 20: Spearman Correlation Test and Spearman's correlation coefficient

	Age	Weight	RunTime	RestPulse	RunPulse	MaxPulse
Age	1.00000	-0.16152 (0.3853)	0.15883 (0.3934)	-0.11766 (0.5285)	-0.29810 (0.1033)	-0.38682 (**0.0316**)
Weight	-0.16152 (0.3853)	1.00000	0.07483 (0.6891)	-0.02958 (0.8745)	0.07541 (0. 6868)	0.14260 (0.4441)

RunTime	0.15883 (0.3934)	0.07483 (0.6891)	1.00000	0.48618 (**0.0056**)	0.28390 (0.1217)	0.20580 (0.2667)
RestPulse	-0.11766 (0.5285)	-0.02958 (0.8745)	0.48618 (**0.0056**)	1.00000	0.36765 **(0.0419)**	0.32634 (0.0732)
RunPulse	-0.29810 (0.1033)	0.07541 (0. 6868)	0.28390 (0.1217)	0.36765 **(0.0419)**	1.00000	0.93152 (**<.0001**)
MaxPulse	-0.38682 (**0.0316**)	0.14260 (0.4441)	0.20580 (0.2667)	0.32634 (0.0732)	0.93152 (**<.0001**)	1.00000

The result is shown as table 20. The Spearman correlation coefficient between Age and MaxPulse is -0.38682 with P value less than 0.05. Thus, it indicates that the null hypothesis that Rho = 0 can be rejected. That is, variable Age and variable MaxPulse have a significant negative correlation relationship, with Spearman correlation coefficient as -0.38682. The Spearman correlation coefficient between RunTime and RestPulse is 0.48618 with P value less than 0.05. Thus, it indicates that the null hypothesis that Rho = 0 can be rejected. That is, variable RunTime and variable RestPulse have a significant positive correlation relationship, with Spearman correlation coefficient as 0.48618. The Spearman correlation coefficient between RunPulse and RestPulse is 0.36765 with P value less than 0.05. Thus, it indicates that the null hypothesis that Rho = 0 can be rejected. That is, variable RunPulse and variable RestPulse have a significant positive correlation relationship, with Spearman correlation coefficient as 0.36765. The Spearman correlation coefficient between RunPulse and MaxPulse is 0.93152 with P value less than 0.05. Thus, it indicates that the null hypothesis that Rho = 0 can be rejected. That is, variable RunPulse and variable MaxPulse have a highly significant positive correlation relationship with Spearman correlation coefficient as 0.93152.

Therefore, we can see that there is a significant correlation relationship between Age and MaxPulse, RunTime and RestPulse, RunPulse and RestPulse, RunPulse and MaxPulse, respectively. Specially, variable RunPulse and variable MaxPulse are highly correlated with each other, since only their correlation coefficients are greater than 0.8 among these significantly correlated paired variables.

III Build a “full model” with exploring VIF

1. Full model with checking VIF

Table 21: Analysis of variance

Source	DF	Sum of squares	Mean Square	F Value	Pr > F
Model	6	722.54361	120.42393	22.43	<0.0001
Error	24	128.83794	5.36825		
Corrected Total	30	851.38154			

Table 22: Statistical indicators

Root MSE	Dependent mean	Coeff Var	R-Square	Adj R-sq
2.31695	47.37581	4.89057	0.8487	0.8108

Table 23: Parameter Estimates with VIF

Variable	DF	Parameter Estimates	Standard Error	t Value	Pr > |t|	VIF
Intercept	1	102.93448	12.40326	8.30	<0.0001	0
Age	1	-0.22697	0.09984	-2.27	0.0322	1.51284
Weight	1	-0.07418	0.05459	-1.36	0.1869	1.15533
RunTime	1	-2.62865	0.38456	-6.84	<0.0001	1.59087
RestPulse	1	-0.02153	0.06605	-0.33	0.7473	1.41559
RunPulse	1	-0.36963	0.11985	-3.08	0.0051	8.43727
MaxPulse	1	0.30322	0.13650	2.22	0.0360	8.74385

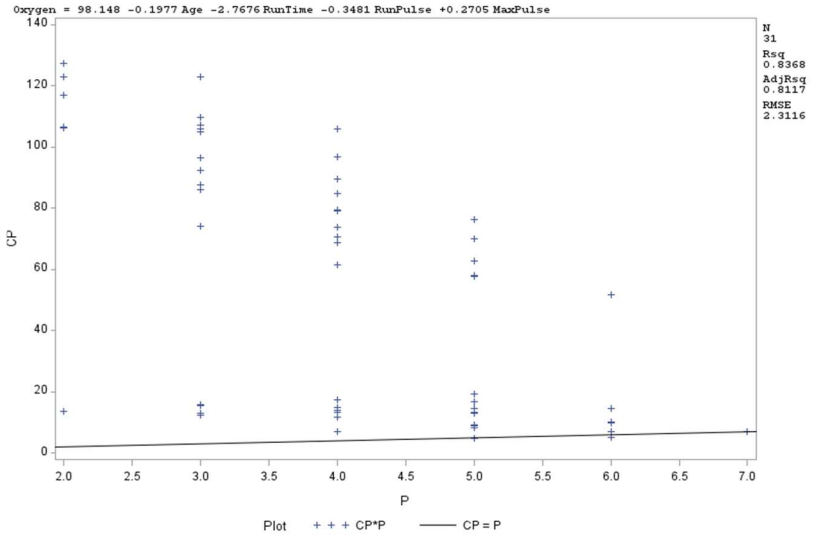
The table 21 shows the p-value of the F test is less than 0.05, which indicates the full model is significant to describe the relationship between variable Oxygen and explanatory variables. Besides, the table 22 shows the R-Square as 0.8487, with adjusted R-Square as 0.8108, which is big enough. These indicate these explanatory variables in full model can explain a big amount of variability of the response variable Oxygen. Then, the table 23 shows the p-value of the t test for the estimated coefficient of explanatory variables. For variable Age, RunTime, RunPulse and MaxPulse, each of them is significant to explain the change of the variable Oxygen when given the remaining variables, since the P-value of each of them is less than 0.05. However, for variable Weight and RestPulse, each of them is not significant to explain the change of the variable Oxygen when given the remaining variables, since the P-value of each of them is greater than 0.05.

Finally, for the VIF, only variable RunPulse and MaxPulse have the bigger value of VIF, greater than 5, which indicates that each of them has a high linear relationship with remaining ones. Therefore, the multicollinearity exists between them, but it may be not that clear since the value of VIF is still lower than 10.

IV Perform all possible subsets regression

1. Explore the plot of Cp vs. p

Figure 1: Cp vs. p



There are 63 all possible subsets regressions, the top three best models regarding cp show as below and the plot of Cp vs. p is shown as above.

Table 24: Top three models

Num. in Model	C(p)	R-Square	Variable in Model
4	4.8800	0.8368	Age RunTime RunPulse MaxPulse
5	5.1063	0.8480	Age Weight RunTime RunPulse MaxPulse
5	6.8461	0.8370	Age RunTime RestPulse RunPulse MaxPulse

2. Identify and build the "best model"

Based on table 24, the best model is containing variable Age, RunTime, RunPulse and MaxPulse. Then we will use these variables to build the model.

Table 25: Analysis of variance

Source	DF	Sum of squares	Mean Square	F Value	Pr > F
Model	4	712.45153	178.11288	33.33	<0.0001
Error	26	138.93002	5.34346		
Corrected Total	30	851.38154			

Table 26: Statistical indicators

Root MSE	Dependent mean	Coeff Var	R-Square	Adj R-sq
2.31159	47.37581	4.87927	0.8368	0.8117

Table 27: Parameter Estimates

Variable	DF	Parameter Estimates	Standard Error	t Value	Pr > |t|
Intercept	1	98.14789	11.78569	8.33	<0.0001
Age	1	-0.19773	0.09564	-2.07	0.0488
RunTime	1	-2.76758	0.34054	-8.13	<0.0001
RunPulse	1	-0.34811	0.11750	-2.96	0.0064
MaxPulse	1	0.27051	0.13362	2.02	0.0533

The table 25 shows the p-value of the F test is less than 0.05, which indicates the best model is significant to describe the relationship between variable Oxygen and explanatory variables. Besides, the table 26 shows the R-Square as 0.8368, with adjusted R-Square as 0.8117, which is big enough. These indicate these explanatory variables in the best model can explain a big amount of variability of the response variable Oxygen. Then, the table 27 shows the p-value of the t test for the estimated coefficient of explanatory variables. For variable Age, RunTime, RunPulse, each of them is significant to explain the change of the variable Oxygen when given the remaining variables, since the P-value of each of them is less than 0.05. However, for variable MaxPulse, we can say that it is not significant to explain the change of the variable Oxygen when given the remaining variables, since the P-value, as

0.0533, is slightly greater than 0.05.

V Perform different stepwise methods

1. Forward selection and the best model of forward selection

Table 28: Summary of forward selection

Step	Variable Entered	Number Var in	Partial R-Square	Model R-Square	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

Table 29: Analysis of variance

Source	DF	Sum of squares	Mean Square	F Value	Pr > F
Model	5	721.97309	144.39462	27.90	<0.0001
Error	25	129.40845	5.17634		
Corrected Total	30	851.38154			

Table 30: Parameter Estimates

Variable	DF	Parameter Estimates	Standard Error	t Value	Pr > |t|
Intercept	1	102.20428	11.97929	8.53	<0.0001
Age	1	-0.21962	0.09550	-2.30	0.0301
Weight	1	-0.07230	0.05331	-1.36	0.1871
RunTime	1	-2.68252	0.34099	-7.87	<0.0001
RunPulse	1	-0.37340	0.11714	-3.19	0.0038
MaxPulse	1	0.30491	0.13394	2.28	0.0316

Table 31: Statistical indicators

C(p)	R-Square	Adj R-Square
5.1063	0.8480	0.8176

For the forward selection, we find the best model contains variable RunTime, RunPulese, Age, Weight and MaxPulse, with C(p) as 5.1063 and Adjusted R-Square as 0.8176. For this model, the whole model is significant and for variables in this model, only Varibale Weight is not significant, since its P-Value is greater than 0.05.

2. Backward selection and the best model of backward selection

Table 32: Summary of backward selection

Step	Variable Removed	Number Var in	Partial R-Square	Model R-Square	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

Table 33: Analysis of variance

Source	DF	Sum of squares	Mean Square	F Value	Pr > F
Model	4	712.45153	178.11288	33.33	<0.0001
Error	26	138.93002	5.34346		
Corrected Total	30	851.38154			

Table 34: Parameter Estimates

Variable	DF	Parameter Estimates	Standard Error	t Value	Pr > |t|
Intercept	1	98.14789	11.78569	8.33	<0.0001
Age	1	-0.19773	0.09564	-2.07	0.0488
RunTime	1	-2.76758	0.34054	-8.13	<0.0001
RunPulse	1	-0.34811	0.11750	-2.96	0.0064
MaxPulse	1	0.27051	0.13362	2.02	0.0533

Table 35: Statistical indicators

C(p)	R-Square	Adj R-Square
4.8800	0.8368	0.8117

For the backward selection, we find the best model contains variable RunTime, RunPulese, Age and MaxPulse, with C(p) as 4.8800 and Adjusted R-Square as 0.8117. For this model, the whole model is significant and for variables in this model, only Varibale MaxPulse is not significant, since its P-Value is slightly greater than 0.05.

3. Stepwise selection and the best model of backward selection

Table 36: Summary of stepwise selection

Step	Variable Entered	Number Var in	Partial R-Square	Model R-Square	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

Table 37: Statistical indicators

C(p)	R-Square	Adj R-Square
4.8800	0.8368	0.8117

Table 38: Analysis of variance

Source	DF	Sum of squares	Mean Square	F Value	Pr > F
Model	4	712.45153	178.11288	33.33	<0.0001
Error	26	138.93002	5.34346		
Corrected Total	30	851.38154			

Table 39: Parameter Estimates

Variable	DF	Parameter Estimates	Standard Error	t Value	Pr > |t|
Intercept	1	98.14789	11.78569	8.33	<0.0001
Age	1	-0.19773	0.09564	-2.07	0.0488
RunTime	1	-2.76758	0.34054	-8.13	<0.0001
RunPulse	1	-0.34811	0.11750	-2.96	0.0064
MaxPulse	1	0.27051	0.13362	2.02	0.0533

For the stepwise selection, we find the best model contains variable RunTime, RunPulese, Age and MaxPulse, with C(p) as 4.8800 and Adjusted R-Square as 0.8117. For this model, the whole model is significant and for variables in this model, only Varibale MaxPulse is not significant, since its P-Value is slightly greater than 0.05.

VI The parsimonious model and check model assumptions

1. Select the best model

Based on previous analysis, we can see that, the best model of C(p) selection, backward selection and stepwise selection are the same. So we will use this model as our parsimonious model which contains variable RunTime, RunPulese, Age and MaxPulse, with C(p) as 4.8800 and Adjusted R-Square as 0.8117.

2. Check model assumptions

i. VIF

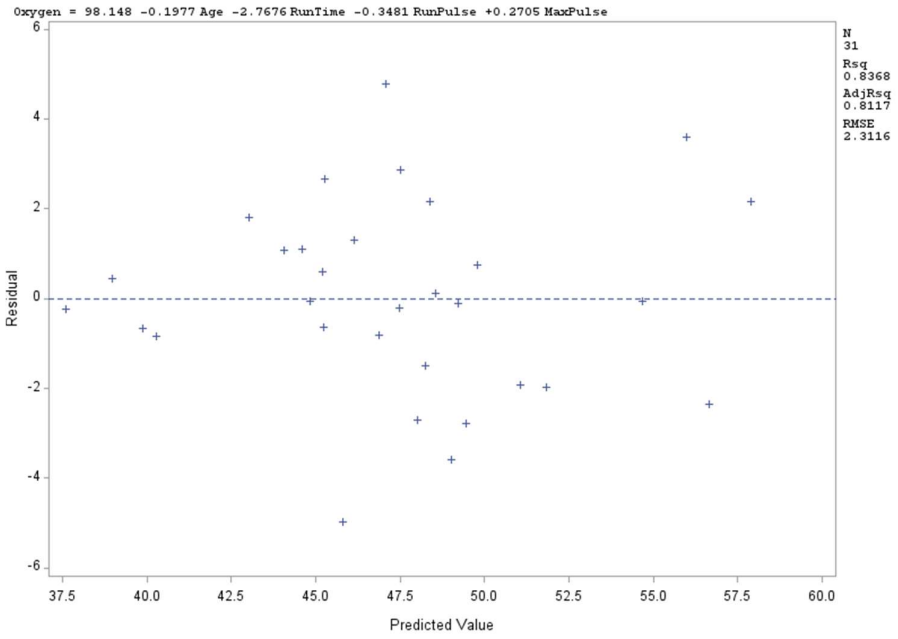
Table 40: Parameter Estimates

Variable	DF	Parameter Estimates	Standard Error	t Value	Pr > |t|	VIF
Intercept	1	98.14789	11.78569	8.33	<0.0001	0
Age	1	-0.19773	0.09564	-2.07	0.0488	1.39464
RunTime	1	-2.76758	0.34054	-8.13	<0.0001	1.25325
RunPulse	1	-0.34811	0.11750	-2.96	0.0064	8.14675
MaxPulse	1	0.27051	0.13362	2.02	0.0533	8.41820

For variable RunPulse and MaxPulse, although the values of VIF are greater than 5 but they are still less than 10. So there is no clear violation for model assumption in this degree.

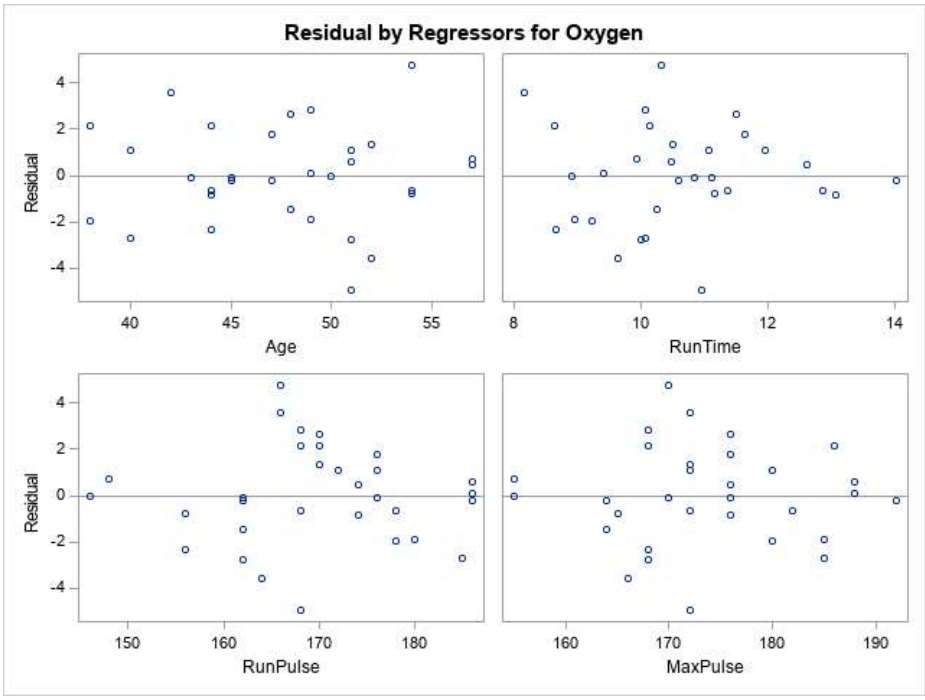
ii. The plot of residuals vs. fitted values

Figure 2: Residuals vs. fitted values



iii. The plot of residuals vs. each explanatory variables

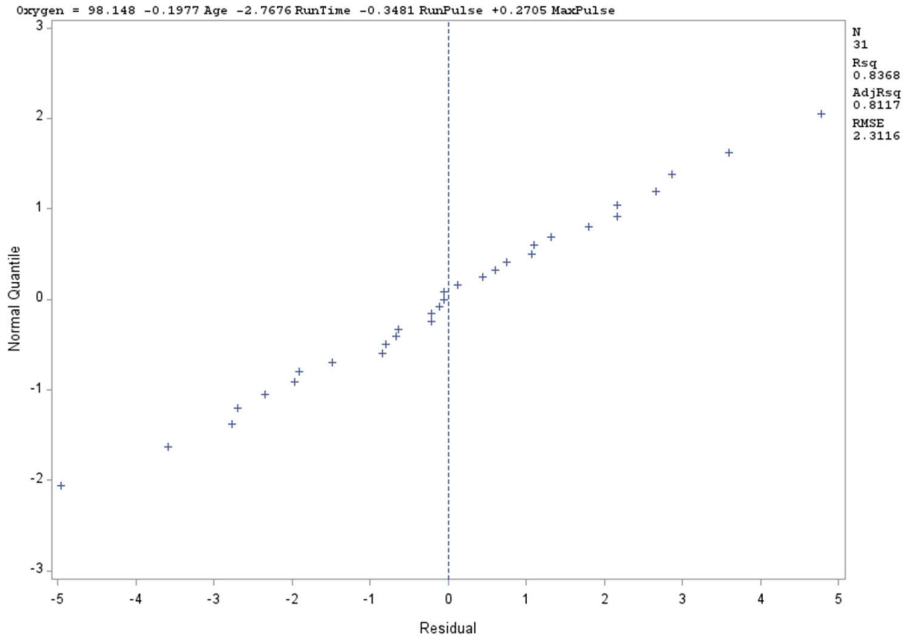
Figure 3: Residuals vs. each explanatory variables



For these residual plots, there is no strange pattern contained in these plots, all plots show horizontal pattern, which indicates the model has the constant variance and this model has detected all variation that contributed by these four variables.

iv. The Q-Q plot for residuals

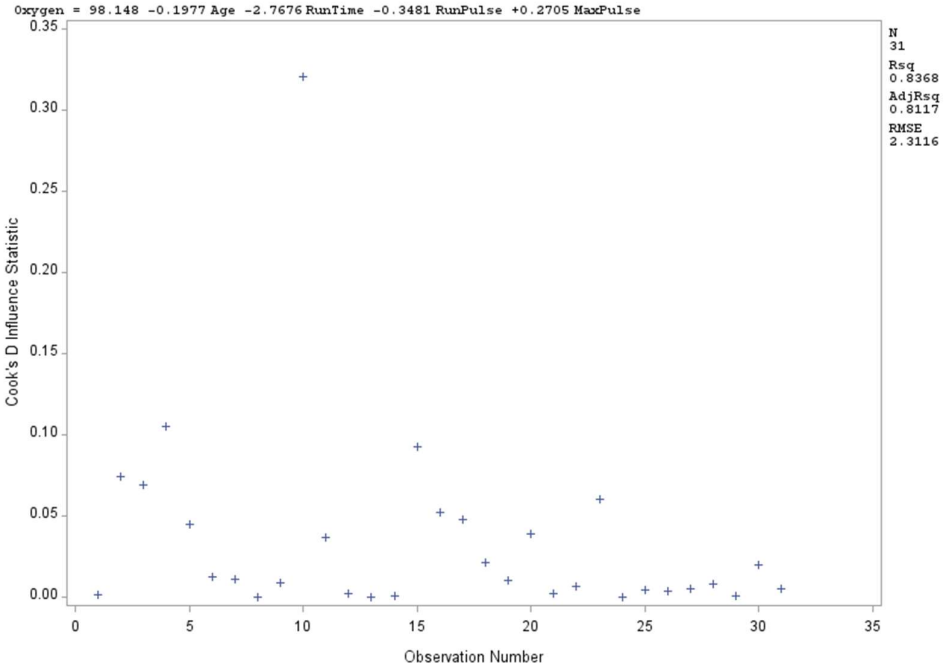
Figure 4: The Q-Q plot for residuals



From this plot we can see, the pattern of Q-Q plot is like a diagonal straight line which indicates that the residuals follow the normal distribution.

v. The plot Cook's distance for all observations

Figure 5: The plot Cook's distance for all observations



From the plot we can see that all values of Cook's distance for all observations are less than 1, which indicates that there is no undue influence on the estimated regression

coefficients for these observations.

VII Conclusion

Based on above analysis, we can conclude that in simple linear regressions, only variable RunTime, RunPulse and RestPulse have significant impacts on the oxygen intake rate; there is a significant correlation relationship between Age and MaxPulse, RunTime and RestPulse, RunPulse and RestPulse, RunPulse and MaxPulse, respectively. Specially, variable RunPulse and variable MaxPulse are highly correlated with each other, since only their correlation coefficients are greater than 0.8 among these significantly correlated paired variables; for the VIF, only variable RunPulse and MaxPulse have the bigger value of VIF, greater than 5, which indicates that each of them has a high linear relationship with remaining ones. Therefore, the multicollinearity exists between them, but it may be not that clear since the value of VIF is still lower than 10; our parsimonious model is the one contains variable RunTime, RunPulese, Age and MaxPulse, with C(p) as 4.8800 and Adjusted R-Square as 0.8117; also, there is no clear violation of our model assumptions for our parsimonious model.

Appendix

```
/*Part A*/
proc import datafile = 'E:/GW/Textbook/Data Analysis/HW5/HW5.csv' /*read the file into
sas*/
dbms = csv /*specify the format of the file*/
out=work.HW5;      /*specify the saved dataset in sas*/
getnames=yes;      /*get the name of the variables from the original file*/
run;    /*run this procedure*/

/* Simple regression */
/* Model 1 : Oxygen ~ Age */
proc reg data = HW5;
model Oxygen = Age;
output out = model1 p = yhat r= mres;
run;
quit;

/* Model 2 : Oxygen ~ Weight */
proc reg data = HW5;
model Oxygen = Weight;
output out = model2 p = yhat r= mres;
run;
quit;

/* Model 3 : Oxygen ~ RunTime */
proc reg data = HW5;
model Oxygen = RunTime;
output out = model3 p = yhat r= mres;
run;
quit;

/* Model 4 : Oxygen ~ RestPulse */
proc reg data = HW5;
model Oxygen = RestPulse;
output out = model4 p = yhat r= mres;
run;
quit;

/* Model 5 : Oxygen ~ RunPulse */
proc reg data = HW5;
model Oxygen = RunPulse;
output out = model5 p = yhat r= mres;
run;
quit;
```

```
/* Model 6 : Oxygen ~ MaxPulse */
proc reg data = HW5;
model Oxygen = MaxPulse;
output out = model6 p = yhat r= mres;
run;
quit;

/* Pearson & Speanman correlation */
proc corr data = HW5 pearson spearman;
var Age Weight Runtime Restpulse Runpulse Maxpulse;
run;

/* Full model & VIF */
proc reg data = HW5;
model Oxygen = Age Weight Runtime Restpulse Runpulse Maxpulse / vif;
output out=Fullmodel p = yhat r = mse;
run;

/* Perform all possible subsets regression and calculate CPs */
proc reg data = HW5;
model Oxygen = Age Weight Runtime Restpulse Runpulse Maxpulse / selection = cp;
plot cp.*np. / cmallows=black;
run;

/* Best model of cp */
proc reg data = HW5;
model Oxygen = Age Runtime Runpulse Maxpulse;
run;

/* Perform different stepwise methods    */
/* Foreward */
proc reg data = HW5;
model Oxygen = Age Weight Runtime Restpulse Runpulse Maxpulse / selection = f
slentry=0.2
                                        slstay=0.1;
run;
/* Best model of forward */
proc reg data = HW5;
model Oxygen = Age Weight Runtime Runpulse Maxpulse;
run;

/* Backward */
proc reg data = HW5;
```

```
model Oxygen = Age Weight Runtime Restpulse Runpulse Maxpulse / selection = b;
run;

/* Best model of backward */
proc reg data = HW5;
model Oxygen = Age Runtime Runpulse Maxpulse;
run;

/* Stepwise */
proc reg data = HW5;
model Oxygen = Age Weight Runtime Restpulse Runpulse Maxpulse / selection = stepwise;
run;

/* Best model of stepwise */
proc reg data = HW5;
model Oxygen = Age Runtime Runpulse Maxpulse;
run;


/* Parsimonious model */
proc reg data = HW5;
model Oxygen = Age Runtime Runpulse Maxpulse/vif;
plot r.*(Age Runtime Runpulse Maxpulse) / nomodel nostat; /* plot residuals vs explanary
variables*/
plot r.*p.; /* plot residuals vs fitted values*/
plot nqq.*r.; /*Q-Q plot of residuals*/
plot cookd.*obs.; /* Cookd of obversions*/
run;
quit;
```