STAT 6210 Homework 5 Report

Xue Ming Wang (Vivian)
G20580112

1. Data Set

The dataset consists of one response variable and six explanatory variables. We aim to investigate the relationship between the response variable, which is the rate of oxygen intake, and the explanatory variables: Age, Weight, RunTime, RestPulse, RunPulse, and MaxPulse. In particular, the RunTime variable indicates the time taken to run 1.5 miles. RestPulse represents the resting heart rate, while RunPulse represents the heart rate during running. MaxPulse represents the maximum heart rate recorded during running. There are 31 observations in total, and we are interested in determining the significant impact of each explanatory variable on the response variable. Additionally, we seek to identify a parsimonious model to explain the oxygen intake rate based on this data set.

2. Explanatory Variables

2.1 Simple Linear Regression

Model	F Value	Pr>F	t value	Pr>t	R-Square
Oxygen = Age	2.97	0.0957	-1.72	0.0957	0.0928
Oxygen = Weight	0.79	0.3817	-0.89	0.3817	0.0265
Oxygen = RunTime	84.01	<.0001	-9.17	<.0001	0.7435
Oxygen = RestPulse	5.50	0.0260	-2.35	0.0260	0.1595
Oxygen = RunPulse	5.46	0.0266	-2.34	0.0266	0.1584
Oxygen = MaxPulse	1.72	0.1997	-1.31	0.1997	0.0560

Table 1. Simple Linear Regression

At first, we start with a simple linear regression with only one explanatory variable. The results indicated that the p-values (Table 1) of Age, Weight, and MaxPulse were greater than 0.05, implying that these variables were not statistically significant when considered individually. However, since the oxygen intake rate is likely dependent on multiple variables, we need to investigate this relationship under more complex conditions.

2.2 Type III SS

Source	DF	Type III SS	Mean Square	F value	Pr>F
Age	1	27.7458	27.7458	5.17	0.0322
Weight	1	9.9106	9.9106	1.85	0.1869
RunTime	1	250.8221	250.8221	46.72	<.0001
RestPulse	1	0.5705	0.5705	0.11	0.7473
RunPulse	1	51.0581	51.0580	9.51	0.0051
MaxPulse	1	26.4914	26.4914	4.93	0.0360

Table 2. Type III SS

Type III sums of squares represent the contribution of each term to a model including all other possible terms. Table 2 displays the outcome of the Type III SS test, which is a crucial method to assess the significance of variables. Through this test, we can determine whether a variable is essential or not when it is introduced into a model that already includes all other variables. According to Table 2, the p-values of the Weight and RestPulse variables are greater than 0.05, indicating that we cannot reject the null hypothesis. This implies that Weight and RestPulse do not have a significant impact on the oxygen intake rate.

2.3 Conclusion

Based on the above analysis, we can draw the conclusion that the variables of Age, RunTime, RunPulse, and MaxPulse are significant factors that would impact the oxygen intake rate. So, these individual explanatory variables are significantly associated with the response.

3. Correlation

As it is commonly known, highly correlated explanatory variables can cause confounding effects. To improve the performance of the model, it is necessary to remove strongly correlated variables. In the following steps, we use the Pearson and Spearman correlation coefficients to examine the relationship between each explanatory variable and use the Variance Inflation Factor (VIF) to assess multicollinearity.

3.1 Pearson and Spearman Correlation

Source	Age	Weight	RunTime	RestPulse	RunPulse	MaxPulse
Age	1	0.2061	0.3092	0.3777	0.0630	0.0150
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	0.0110	1	0.0518	0.0951
RunPulse	0.0630	0.3284	0.0858	0.0518	1	<.0001
MaxPulse	0.0150	0.1761	0.2213	0.0951	<.0001	1

Table 3. P-Value for Pearson 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

Table 4. P-Value for Spearman Correlation Coefficients

Firstly, we evaluated the correlation between each explanatory variable using the Pearson correlation coefficient, and the results are presented in Table 3. To avoid the errors arising from a non-normal distribution, we have also computed the results using the Spearman correlation coefficient, which are presented in Table 4. The p-value of each correlation coefficient is provided

in both tables, and a p-value less than 0.05 indicates a linear relationship between the two variables. From these tables, we can conclude that Age and MaxPulse have a linear relationship, RunTime and RunPulse have a linear relationship, and RunPulse and MaxPulse have a strong linear relationship.

3.2 Multicollinearity

Multicollinearity occurs when two or more explanatory variables in a multiple regression model are highly correlated. To check for multicollinearity, we typically use the VIF indicator. The VIF of an explanatory variable indicates the strength of the linear relationship between the variable and the remaining ones. Generally, if the VIF is greater than 10, we consider the variable to be strongly related to the other variables and identify them as multicollinearity. Table 5 shows the results of the VIF analysis. We can see that none of the variables have a VIF greater than 10, which means that there is no exact variable that is highly correlated with the other variables.

Variable	DF	Parameter Estimate	SE	t value	Pr>t	Variable Inflation
Intercept	1	102.9345	12.4033	8.30	<.0001	0
Age	1	-0.2270	0.0998	-2.27	0.0322	1.5128
Weight	1	-0.0742	0.0546	-1.36	0.1869	1.1553
RunTime	1	-2.6287	0.3846	-6.84	<.0001	1.5909
RestPulse	1	-0.0215	0.0661	-0.33	0.7473	1.4156
RunPulse	1	-0.3696	0.1199	-3.08	0.0051	8.4373
MaxPulse	1	0.3032	0.1365	2.22	0.0360	8.7439

Table 5. VIF Indicator

3.3 Conclusion

Based on the results of the correlation analysis, it can be concluded that there is a linear relationship between Age and MaxPulse, RunTime and RunPulse, and a strong linear relationship between RunPulse and MaxPulse. Furthermore, VIF analysis indicated that there is no clear multicollinearity among the explanatory variables in the full model. Therefore, we can conclude that there is no high correlation between any pair of explanatory variables, and multicollinearity is not a concern in this dataset.

4. Parsimonious model

4.1 Possible Subset Regression

After analyzing the significance and correlation of the explanatory variables, we found that there are four significant variables and no multicollinearity. The next step is to find the parsimonious model. One approach is to generate a list of all possible subsets of the parsimonious model and evaluate them using indicators such as Mallow's statistic C_p and R-Squares. C_p is SSE(reduced)/MSE(full)-(n-2a), where a is the number of parameters in the reduced model. It is a measure of the predictive power of a linear regression model, and a value close to the number of parameters in the model, including the intercept, indicates a good fit. When searching for the best model, it is recommended to consider models with $C_p \leq p$, and to choose the first model with the

fewest variables where C_p approaches p. This approach can help to select a parsimonious model that balances goodness of fit with simplicity. We aim to select the smallest p when C_p equals p.

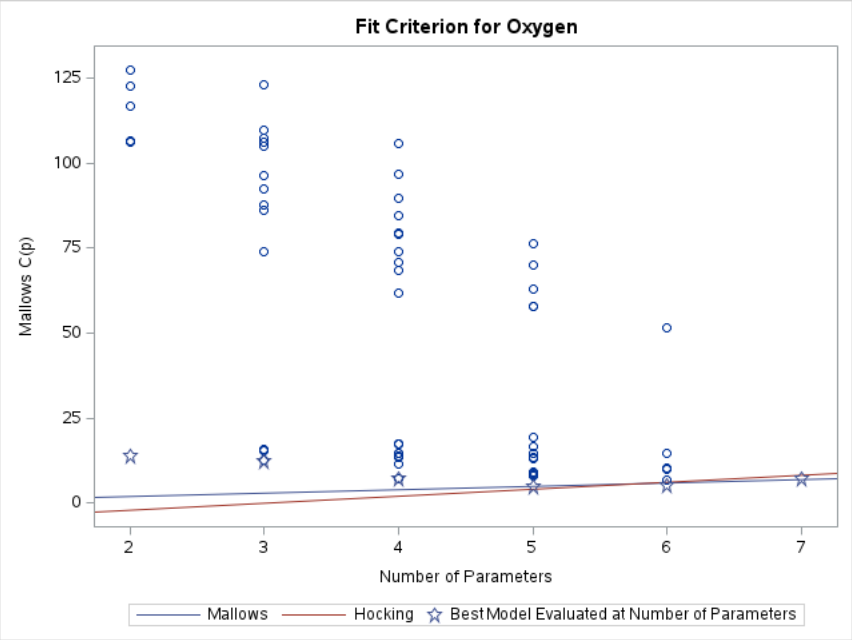

Figure 1. Scatter Plot of C_p vs. P

Figure 1 displays all the possible models' C_p and p, and we seek to identify the point where C_p equals p. Based on this graph, we observe that the line intersects with the C_p *p point at p=5 and p=7. Since we needed to find the smallest C_p, the best model is the one where C_p equals 5. When C_p =5.1063, the variables in the best model are Age, Weight, RunTime, RunPulse, and MaxPulse, where the R-Square for this model is 0.8480, indicating a strong relationship.

4.2 Stepwise Methods

4.2.1 Forward Selection

Let's consider a different approach for selecting variables, known as forward selection. With this method, variables are added to the model one at a time. At each step, any variable not already in the model is evaluated for potential inclusion. The most significant of these variables is added to the model, provided its p-value falls below a predetermined threshold. The process continues until no additional variables meet the inclusion criteria. The resulting model is considered parsimonious. However, it's important to note that the use of multiple hypothesis testing in forward selection can impact the true Type I SS for a variable.

Step	Entered	Number Vars In	Model R^2	F value	Pr>F
1	RunTime	1	0.7434	84.01	<.0001
2	Age	2	0.7642	2.48	0.1267
3	RunPulse	3	0.8111	6.70	0.0154
4	MaxPulse	4	0.8368	4.10	0.0533
5	Weight	5	0.8480	1.84	0.1871

Table 6. Summary of Forward Selection

Table 6 summarizes the results of the forward selection process, which started with the variable RunTime and then added Age, RunPulse, MaxPulse, and Weight. Nonetheless, the order in which variables are added in forward selection can influence the outcome, like a Type I SS table. Therefore, another selection method is needed to identify the best parsimonious model.

4.2.2 Backward Selection

Forward selection has its limitations, as adding a new variable at each step may cause one or more previously included variables to lose their significance. An alternative method that overcomes this issue is backward selection. It involves starting with a model that includes all the relevant variables and gradually removing variables from the model one at a time based on their significance. At each step, the least significant variable is removed from the model as long as its p-value is above a certain threshold. The process is repeated by fitting new reduced models and removing the least significant variable until all remaining variables are statistically significant. The resulting model is considered parsimonious.

Step	Removed	Number Vars In	Model R^2	F value	Pr>F
1	RestPulse	5	0.8480	0.11	0.7473
2	Weight	4	0.8368	1.84	0.1871

Table 7. Summary of Backward Selection

By examining Table 7, we can observe that the backward selection approach removed the RestPulse and Weight variables, which were found to be non-significant as their p-values wer greater than 0.05.

4.2.3 Stepwise Regression

Stepwise selection is a variable selection method that allows for both adding and dropping variables at various steps. Backward stepwise selection begins with a model that includes all candidate variables and then drops variables that are not statistically significant. However, unlike forward selection, backward stepwise selection allows for the reintroduction of previously dropped variables if they later become significant. The process involves selecting the least significant variable to drop and then considering all previously dropped variables (except for the most recently dropped one) for reintroduction into the model. As a result, two separate significance levels must be chosen for variable deletion and variable addition. The significance level for adding variables must be more stringent than the level for deleting variables.

Step	Entered	Removed	Number Vars In	Model R2	F value	Pr>F
1	RunTime		1	0.7434	84.01	<.0001
2	Age		2	0.7642	2.48	0.1267
3	RunPulse		3	0.8111	6.70	0.0154
4	MaxPulse		4	0.8368	4.10	0.0533

Table 8. Summary of Stepwise Selection

According to Table 8, which displays the results of stepwise selection, the variables RunTime, Age, RunPulse, and MaxPulse were added to the model, and no variables were removed.

4.3 Conclusion

Variable	Parameter Estimate	t value	P > t
Intercept	98.1479	8.33	<.0001
Age	-0.1977	-2.07	0.0488
RunTime	-2.7676	-8.13	<.0001
RunPulse	-0.3481	-2.96	0.0064
MaxPulse	0.2705	2.02	0.0533

Table 9: Parameter Estimates

Source	DF	F value	P > f
Model	4	33.33	<.0001

Table 10: Analysis of Variance

Based on the analysis of significant variables and the different selection methods, it can be concluded that the most efficient model includes four variables: Age, RunTime, RunPulse, and MaxPulse. The parsimonious model equation is: Oxygen = 98.1479 - 0.1977Age - 2.7376 RunTime - 0.3481RunPulse + 0.2705MaxPulse (Table 9). The model is significant with a p-value smaller than 0.0001 (Table 10), and each variable in the model is significant with a p-value smaller than 0.05 (Table 9). The R-Square value of 0.8368 indicates that approximately 84.68% of the variation in the response variables is explained by the explanatory variables in the model. So, this is the best parsimonious model built based on the analysis above.

5. Model Assumptions

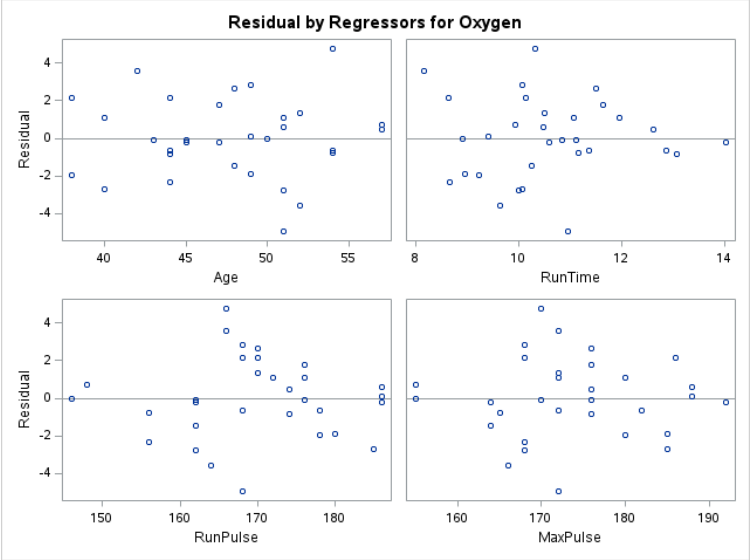

Figure 2: Residual Plot of Oxgyen

After identifying the best parsimonious model, it is crucial to confirm several assumptions. The first assumption is independence. It is important to ensure that the error term is independent for each explanatory variable and the response variable (y-hat). Figure 2 displays the scatter plot of residuals against each explanatory variable, which indicates that there is no discernible pattern between the residuals and the variables. The mean of the residuals is zero and the residuals form a horizontal line, with the dots of the explanatory variable randomly dispersed above or below the residual line. Based on these observations, we can conclude that the error is independent across the explanatory variables, variance is constant, and the model is correct.

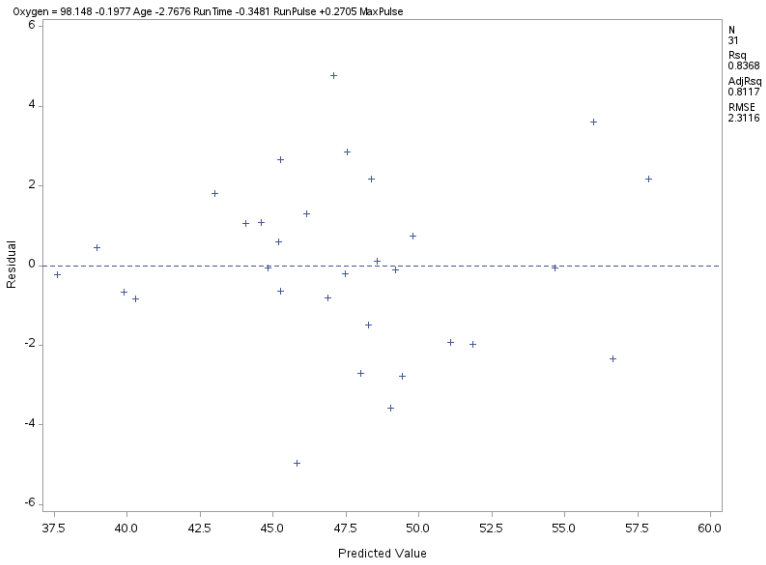

Figure 3: Residual Plot Predict

Next, we need to examine the relationship between the errors and the response variables. Figure 3 shows the plot of errors against predicted values of the response variable. Similar to the previous analysis, the plot reveals a horizontal pattern, with the predicted values randomly scattered above and below the errors. Therefore, we can again conclude that the residuals are independent of the fitted values, variance is constant, and the model is correct.

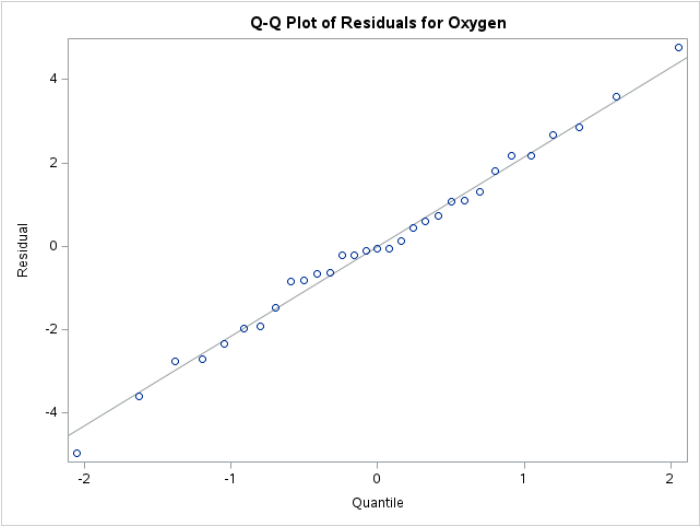

Figure 4. QQ Plot - Normality of Residual

Thirdly, we need to check the assumption of normality of the residuals. Figure 4 shows the QQ plot of the residuals, where a straight line indicates a normal distribution. The plot clearly shows a linear pattern, which means the residuals are normally distributed.

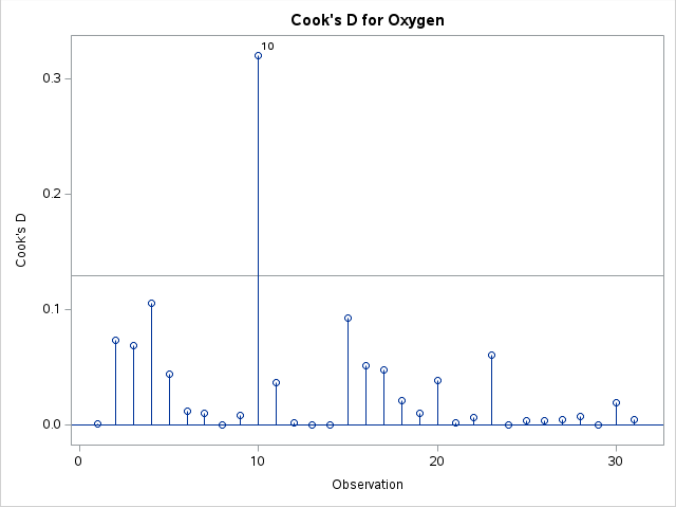

Figure 5: Cook's Distance

Finally, we need to investigate whether there are any outliers. Figure 5 shows the Cook's distance plot, which is a measure of the influence of each observation on the regression model. A large Cook's distance indicates that an observation has a significant impact on the model and is a potential outlier. In this plot, we observe one observation with a larger Cook's distance than any other observations. Therefore, if we aim for a more precise and accurate result, it would be necessary to remove observation 10 with a significantly larger Cook's distance from the dataset.

In summary, the analysis found that the error is independent across the explanatory variables, and the residuals are independent of the fitted values. The residuals are also found to be normally distributed. Additionally, one observation is identified as an outlier, which is recommended to remove. Based on the analysis above, there is no clear violation of the model assumptions, except for a minor outlier to remove.

6. Conclusion

The above analysis indicates that the variables Age, RunTime, RunPulse, and MaxPulse have a significant impact on the oxygen intake rate, and there are linear relationships between Age and MaxPulse, RunTime and RunPulse, as well as a strong linear relationship between RunPulse and MaxPulse. Moreover, the VIF analysis shows no multicollinearity among the explanatory variables in the full model. The parsimonious model equation is Oxygen = 98.1479 - 0.1977Age - 2.7376 RunTime - 0.3481RunPulse + 0.2705MaxPulse. Overall, the model assumptions are not violated, except for a minor outlier that needs to be removed for a more accurate result.