IE5318 - Multiple Linear Regression

By:

Preetam Kulkarni and Mohammadamin Soltanianfard

I. Initial Proposal

Problem description

The shear behavior at the interface of infrastructures and bedrock, especially in concrete dams, plays an essential role in determining the response of structure to horizontal loading. The type of loading can be categorized as monotonic or cyclic such as 1) hydrostatic pressure resulting from the reservoir, 2) earthquake loading, and 3) flood. Knowing this, Soltanianfard et al. (2020) performed a series of laboratory and field tests in an experimental program to investigate the interface behavior of rock and concrete. Two distinct types of concretes were used in the study which are: 1) Conventional vibrated concrete (CVC) and 2) Roller compacted concrete (RCC). They considered three different concrete uniaxial strengths for each series of tested specimens. Also, different values of normal stress were considered for performing the shear tests. Different values of maximum shear force were required to break the interfaces due to various bonding conditions. The resulting shear stress - shear displacement curves were presented and discussed in the research paper. But in that study, the variation of shear strength with different variables presented in this proposal was not investigated. Since the shear strength at the interface is usually considered as the failure stress, which is usually used to determine the sliding safety factor (SF) of structures, having the corresponding knowledge being studied in the current project would be vital. The relationship between the shear strength and different variables presented in this proposal, achieved through an experimental investigation, would help the designers to predict the behavior of concrete dam structures with more reliable knowledge.

Variables

In this study, the quantitative response variable is the "Shear strength".
Predictor variables are as follows:

1. Normal stress – Stress perpendicular to the shear surface at the point where max. shear stress occurs
2. Maximum shear force – Peak value of shear force at which shear strength is defined
3. Uniaxial compressive strength – The strength at which the specimen fails under uniaxial compressive force
4. Type of specimen being tested at the interface (Binary Variable)
    a. Rock-CVC (1)
    b. Rock-RCC (0)

Discussion on the matrix scatter plot and correlation matrix of the variables

The scatter plots shown on the next page in a matrix form were generated in SAS. The plot has both Response vs. Predictors and Predictor vs. Predictors. In our discussion, we will refer to the plots which are above the diagonal of matrix scatterplots. We will also be looking at the correlation values from Figure 2.

Shear Strength vs. predictors

- Shear strength is strongly correlated with Max shear Force at 0.99
- Shear strength increases with increase in normal stress and has a correlation which is not high
- There is an overall downward trend in shear strength vs. UCS plot as we move from CVC to RCC
    - However, within the categories of CVC and RCC, there is an upward trend
    - There can be an interaction between UCS and Interface type

    - The correlation of -0.208 between UCS and shear strength is not high
- Shear strength appears to be higher for CVC than RCC and the correlation between shear strength and Interface type is not high

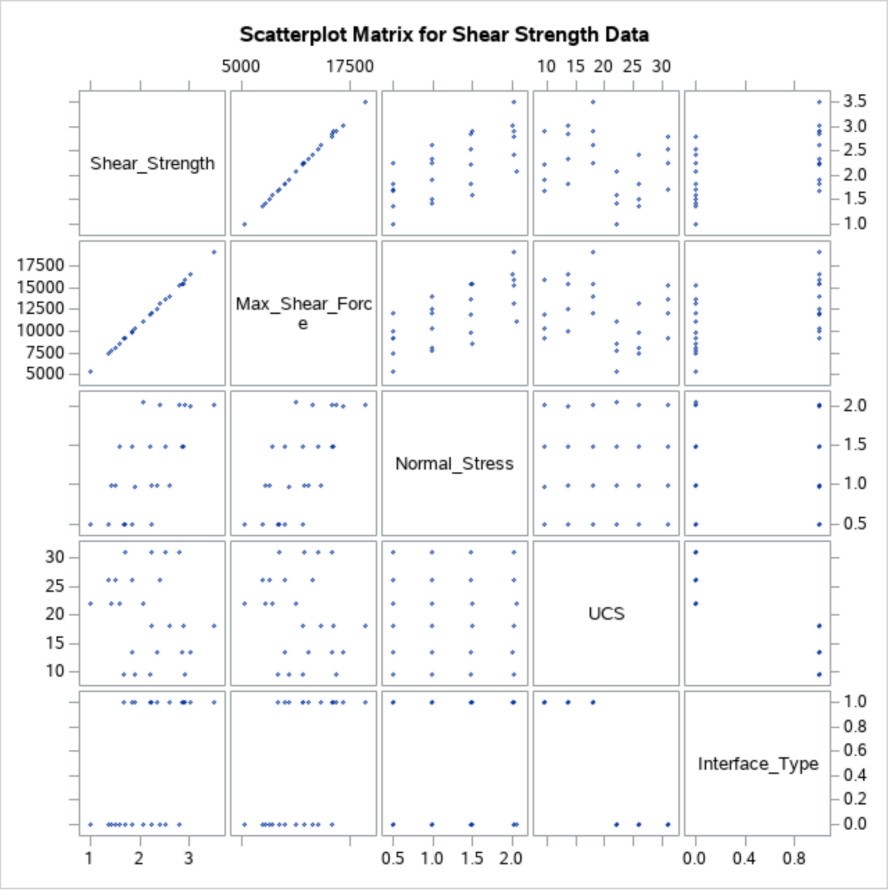

Figure 1 Scatterplot Matrix - Initial proposal

Predictor vs. Predictors

- Max. shear force increases with normal stress and correlation is close to being flagged as high
- Max. shear force is negatively correlated with UCS
    - Within the categories of CVC and RCC, there is an upward trend
    - The correlation is weak at -0.207
- Max. shear force is higher for CVC when compared to RCC
    - Correlation is not high between Max. shear force and Interface type
- Normal stress in uncorrelated with UCS which reflects in the very low correlation value of 0.0031
- Normal stress and Interface type do not show any trend and the correlation value is -0.0051
- UCS and interface type predictors appear to have a problem because of high correlation of -0.87
    - UCS for RCC is higher when compared to that of CVC

Overall, it is good that none of the "y vs. predictor" plots show curvature and they all appear to have a linear relationship. There are also no outliers in any of the plots.

Pearson Correlation Coefficients, N = 24					
	Shear_Strength	Max_Shear_Force	Normal_Stress	UCS	Interface_Type
Shear_Strength	1.00000	0.99963	0.68410	-0.20859	0.51310
Max_Shear_Force	0.99963	1.00000	0.68961	-0.20728	0.50845
Normal_Stress	0.68410	0.68961	1.00000	0.00315	-0.00510
UCS	-0.20859	-0.20728	0.00315	1.00000	-0.87063
Interface_Type	0.51310	0.50845	-0.00510	-0.87063	1.00000

Figure 2 Correlation matrix - Initial proposal

Potential complications

As discussed earlier, the high correlation between UCS and Interface type may create problems such as high variance inflation factor.

Max shear force vs. Normal stress is also close to being highly correlated, however, further analysis of VIF and evaluating adding interaction terms instead of individual predictors is required.

I. New Proposal

As the data collected by us initially had very high multicollinearity and VIFs. Another issue with the data set was that the predictors were mathematically related. Hence, it was necessary to use a different dataset after consulting with Dr. Chen. Below is a discussion on the new dataset.

Problem description

A person suffering from cardiovascular disease, specially if they are older, are required to take necessary care while exercising. In these cases, when the heartbeat increases beyond a certain level, complications may arise. A person might get exhausted easily if the amount of exercise is beyond his/her stamina. The intensity of workout can be determined by the heartrate. Therefore, fitness trainers are required to consider heartrate while proposing workout programs to their trainees.

Variables

The response variable for the current project is "Heartrate" measured in beats per minute. The data was collected from people performing cycling at the Maverick Activities Center. Below are the predictors:

- RPM (x_1)– Measure of output of the cycling machine measured in revolutions per minute
- Incline level (x_2) – Measures of difficulty level for the exercise
- Weight (x_3) of a person performing the exercise
- Age (x_4) of a person performing the exercise

Heartbeat(Y)	***RPM(x_1)***	***Level(x_2)***	***Weight(x_3)***	***Age(x_4)***
139	57	8	130	23
125	60	5	110	22
139	77	3	161	23
111	77	2	170	21
98	73	7	140	35
133	116	1	190	23
137	60	1	90	18
152	86	14	220	22
138	76	6	118	24
159	93	3	105	22
122	73	4	135	24
139	96	12	163	21
137	55	15	215	42
128	73	8	173	26
147	98	6	198	25
121	70	15	198	23
114	83	16	210	43
140	86	2	205	18

117	120	1	175	25
136	80	5	181	24
139	81	14	154	23
123	74	9	140	17
132	90	7	210	27
133	66	11	143	23
104	90	8	215	23
128	80	10	172	24
91	68	2	195	56
110	70	5	145	20
151	83	18	154	22
82	83	1	200	52
95	84	2	165	18
88	80	2	190	60
121	81	4	185	30
91	82	3	205	57
93	79	3	140	35

Table 1 Data set for the New Proposal

Discussion on the matrix scatter plot and correlation matrix of the variables

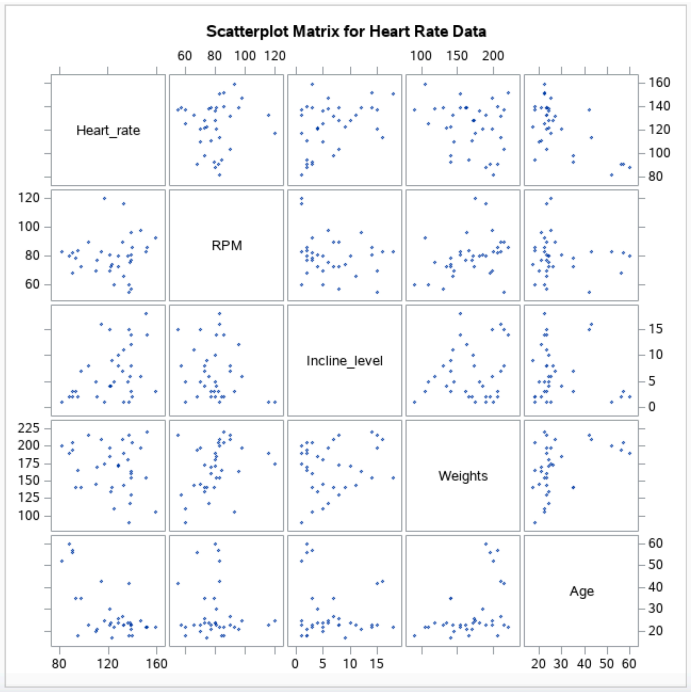

Figure 3 Scatterplot Matrix - New proposal

The scatter plots shown above in a matrix form were generated in SAS. The plot has both Response vs. Predictors and Predictor vs. Predictors plots. In our discussion, we will refer to the plots which are above the diagonal of matrix scatterplots. We will also be looking at the correlation values from Figure 4.

Heartrate vs. predictors

- Heartrate does not appear to be well correlated with RPM
  - Correlation of 0.0708 is low
  - There are two possible x-outliers near the value of 120 RPM
- Heartrate vs. incline level shows an upward trend
  - The values are more spread-out
  - The correlation of 0.398 is not high
- Heartrate vs. weight plot shows a downward trend
  - The correlation is not be high because the values are spread-out
- Heartrate vs. Age plot shows a downward trend
  - There could be slight curvilinear relationship but it is probably weak
  - The correlation is strong because it is -0.641

Pearson Correlation Coefficients, N = 35					
	Heart_rate	RPM	Incline_level	Weights	Age
Heart_rate	1.00000	0.07086	0.39873	-0.21636	-0.64123
RPM	0.07086	1.00000	-0.20532	0.33651	-0.09480
Incline_level	0.39873	-0.20532	1.00000	0.18414	-0.14254
Weights	-0.21636	0.33651	0.18414	1.00000	0.40988
Age	-0.64123	-0.09480	-0.14254	0.40988	1.00000

Figure 4 Correlation matrix – New proposal

Predictor vs. Predictors

- RPM vs. incline level plot shows a downward trend and correlation is not high
- RPM vs. weight plot shows an upward trend and the correlation of 0.336 is not high
- RPM doesn't decrease much with increase in age and the correlation is weak
- All the RPM vs. predictor plots show two possible x-outliers near the value of 120 RPM
- Incline level vs. weight plot shows an upward trend
  - The correlation is weak because the values are spread-out
- Incline level vs. age plot shows a downward trend and the correlation is weak
  - There are four possible x-outliers near the age values of 50 to 60 years
- Weight vs. age shows an upward trend and the correlation of 0.4098 is not high
  - There are four possible x-outliers near the age values of 50 to 60 years

Overall, it is good that none of the "y vs. predictor" plots show curvature and they all appear to have a linear relationship. There could be a few x-outliers among the values of RPM and Age. However, they might be appearing as an outlier in the scatterplots because of lack of data points as well. This can only be confirmed after further analysis of residual plots and outlier tests.

Our conclusions from the matrix scatterplots and the correlation matrix with regards to multicollinearity is that it exists but it might not be a serious problem because all the correlation values between predictors are less than 0.7.

Potential complications

Outliers are a possibility; however, this cannot be confirmed until further analysis of residual plots and outlier test are conducted. Curvilinearity and multicollinearity don't seem to be a problem.

II. Preliminary Multiple Linear Regression Model Analysis

The SAS output for the fitted preliminary model is as follows:

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	7298.42182	1824.60545	8.18	0.0001
Error	30	6693.74961	223.12499		
Corrected Total	34	13992			

Root MSE	14.93737	R-Square	0.5216
Dependent Mean	123.22857	Adj R-Sq	0.4578
Coeff Var	12.12168		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Type I SS	Variance Inflation
Intercept	1	134.42796	18.55466	7.24	<.0001	531485	0
RPM	1	0.19601	0.21799	0.90	0.3757	70.26430	1.42226
Incline_level	1	1.51340	0.58235	2.60	0.0144	2495.00474	1.28566
Weights	1	-0.06654	0.09733	-0.68	0.4995	2084.41260	1.73582
Age	1	-0.90904	0.26384	-3.45	0.0017	2648.74018	1.47678

Figure 5 SAS output for Preliminary model with four predictors

The fitted model is given by the following equation:

$$\hat{y}_i = 134.428 + 0.196\, x_{i1} + 1.513\, x_{i2} - 0.066\, x_{i3} - 0.909\, x_{i4}$$

Model assumptions

Model form

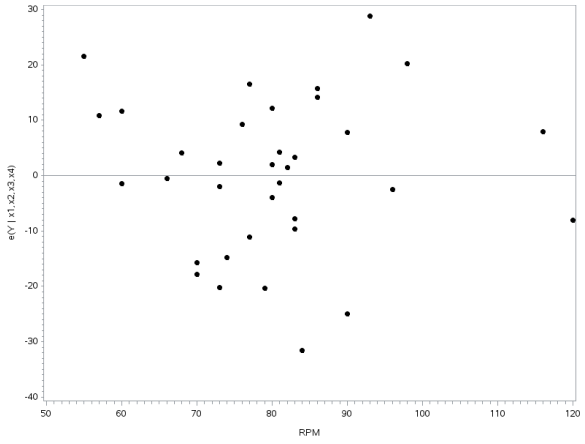

Figure 6 Residual vs. RPM (x_1)

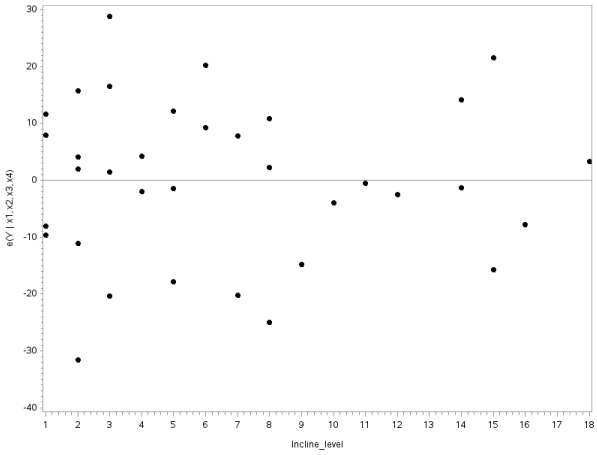

Figure 7 Residual vs. Incline level (x_2)

Figures 6 – 9 are the "Residual vs. Predictor" plots. These plots do not show any curvature. It can be seen in *Figure 9* that the residuals are clustered between age group of 20 and 30. There is a possibility of outliers between age group of 50 and 60 as noticed in the scatterplots. Similarly, even in *Figure 6*, there is a

possibility of outliers between the RPM values of 110 and 120. As the "*y* vs. Predictors" plots also did not show any curvature, hence, transformations are not required. Also, MLR model form is adequate.

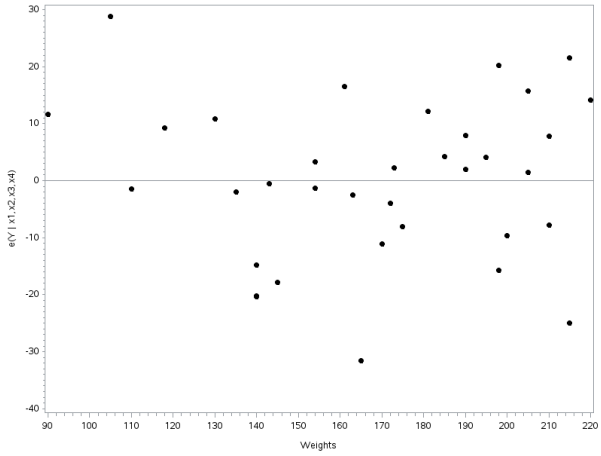

Figure 8 Residual vs. Weight (x_3)

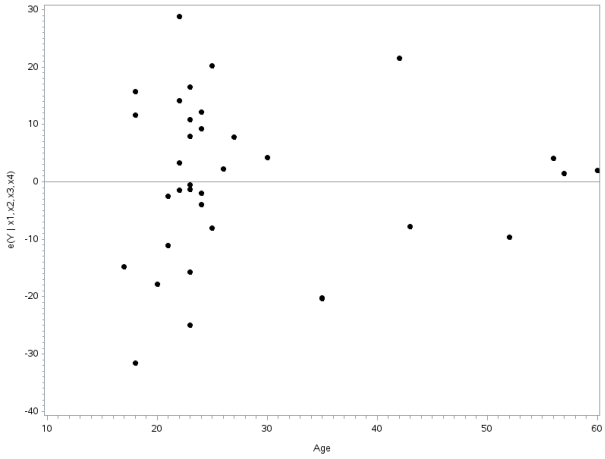

Figure 9 Residual vs. Age (x_4)

Constant variance

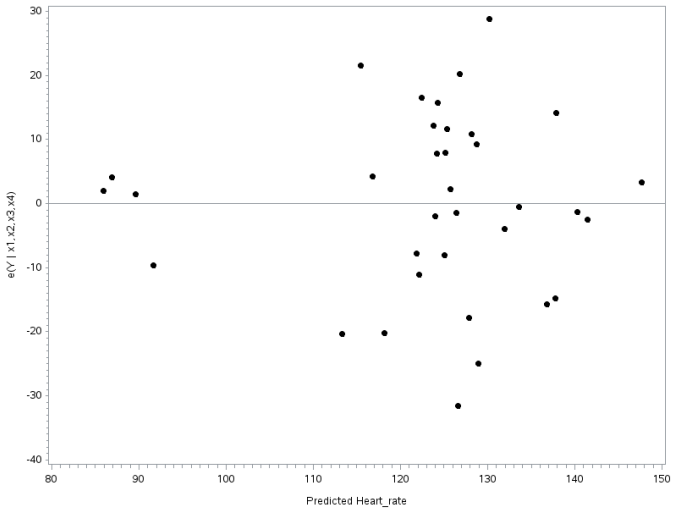

Figure 10 Residual vs. Predicted Heartrate $(\hat{y})$

Figure 10 shows the "*e* vs. $\hat{y}$" plot. It can be seen that the points are randomly scattered. The points are clustered beyond heartrate of 110 beats per minute. There is no funnel and hence, variance appears to be constant. If there were more data points between the heartrates of 90 and 110, we might see a funnel shape. Curvature is also not seen in this plot. No observations appear to be an outlier in this plot though the matrix scatterplots showed possibility of outliers.

Normality

The normal probability plot in *Figure 11* shows that a straight line fits the points and normality appears to be satisfied. No outliers seen in this plot as well.

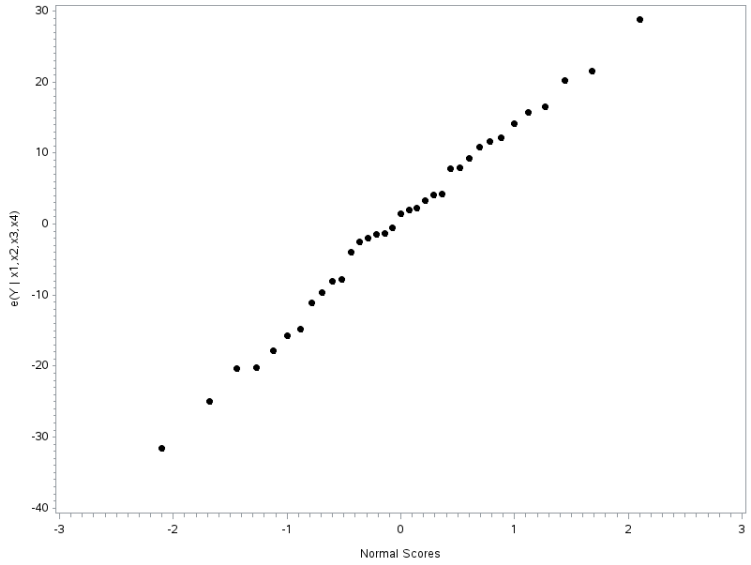

Figure 11 Normal probability plot

Diagnostics

Observation #	Response	Predictors				Residual	RStudent	cookdi	hii	DFFITS	DFBETAS				
	Heart_rate	RPM	Incline_level	Weights	Age						Intercept	RPM	Incline_level	Weights	Age
1	139	57	8	130	23	10.8501	0.7701	0.01675	0.1223	0.2875	0.2512	-0.1838	-0.0048	-0.043	-0.0509
2	125	60	5	110	22	-1.4376	-0.1021	0.00035	0.1407	-0.0413	-0.0368	0.0182	0.0065	0.0163	0.0037
3	139	77	3	161	23	16.5595	1.1535	0.01867	0.0662	0.3072	0.1685	-0.1284	-0.2052	0.0976	-0.1585
4	111	77	2	170	21	-11.1464	-0.7859	0.01544	0.1099	-0.2761	-0.1336	0.1374	0.2058	-0.1462	0.178
5	98	73	7	140	35	-20.1988	-1.4393	0.03756	0.0858	-0.441	-0.1083	-0.0603	-0.1375	0.3079	-0.2811
6	133	116	1	190	23	7.8714	0.5968	0.02263	0.2371	0.3327	-0.1795	0.2301	-0.0754	0.0108	-0.0368
7	137	60	1	90	18	11.6491	0.8854	0.04711	0.2298	0.4836	0.4303	-0.1863	-0.1887	-0.1776	-0.088
8	152	86	14	220	22	14.1646	1.0396	0.04286	0.1658	0.4635	-0.1566	-0.0153	0.1646	0.2682	-0.1847
9	138	76	6	118	24	9.2632	0.647	0.0094	0.0992	0.2147	0.092	0.0476	0.0405	-0.1692	0.0498
10	159	93	3	105	22	28.7882	2.3477	0.2778	0.2248	1.2641	0.0466	0.8085	0.1568	-1.0428	0.3581
11	122	73	4	135	24	-1.9908	-0.1355	0.00026	0.0636	-0.0353	-0.0255	0.0083	0.0112	0.0117	0.0043
12	139	96	12	163	21	-2.4703	-0.1765	0.00114	0.1501	-0.0742	0.0326	-0.0543	-0.051	0.0322	-0.0094
13	137	55	15	215	42	21.5758	1.746	0.20986	0.2689	1.0588	0.121	-0.6225	0.3025	0.4542	0.1147
14	128	73	8	173	26	2.3021	0.1551	0.00023	0.0448	0.0336	0.0142	-0.0177	-0.002	0.0129	-0.012
15	147	98	6	198	25	20.183	1.4404	0.03892	0.0886	0.449	-0.2388	0.2035	-0.0364	0.1535	-0.114
16	121	70	15	198	23	-15.7673	-1.1519	0.04673	0.1511	-0.486	-0.0334	0.1919	-0.1991	-0.2305	0.1733
17	114	83	16	210	43	-7.8496	-0.5891	0.02019	0.2216	-0.3143	0.1681	-0.0945	-0.2415	0.0277	-0.1705
18	140	86	2	205	18	15.6912	1.1974	0.07937	0.2192	0.6345	0.0803	-0.2275	-0.4177	0.4964	-0.4645
19	117	120	1	175	25	-8.0926	-0.6387	0.0348	0.2948	-0.413	0.233	-0.3479	0.0276	0.0994	-0.0427
20	136	80	5	181	24	12.1844	0.8369	0.00896	0.0595	0.2105	0.054	-0.0805	-0.1071	0.1284	-0.1233
21	139	81	14	154	23	-1.3377	-0.0942	0.00026	0.1257	-0.0357	0.0045	-0.0138	-0.03	0.0169	-0.0064
22	123	74	9	140	17	-14.7844	-1.0273	0.01587	0.07	-0.2819	-0.1513	0.0346	-0.0681	0.0748	0.1051
23	132	90	7	210	27	7.8542	0.5436	0.00573	0.0865	0.1672	-0.0614	0.0037	-0.0278	0.1158	-0.0638
24	133	66	11	143	23	-0.5893	-0.0405	0.00003	0.0838	-0.0123	-0.0068	0.0038	-0.0054	0.0034	0.0008
25	104	90	8	215	23	-24.9627	-1.8509	0.08553	0.1189	-0.6799	0.1991	0.0415	0.0942	-0.5106	0.3686
26	128	80	10	172	24	-3.9814	-0.2684	0.00069	0.0444	-0.0579	-0.0003	-0.0028	-0.026	-0.003	0.0118
27	91	68	2	195	56	4.0977	0.3067	0.00561	0.2242	0.1649	0.0099	-0.0526	-0.0514	0.0235	0.0931
28	110	70	5	145	20	-17.887	-1.256	0.02455	0.0735	-0.3537	-0.283	0.1919	0.1421	-0.0342	0.1817
29	151	83	18	154	22	3.3076	0.2518	0.00439	0.2509	0.1458	-0.0333	0.0636	0.1334	-0.0686	0.0322
30	82	83	1	200	52	-9.6326	-0.7027	0.02088	0.1721	-0.3204	0.0737	-0.0081	0.109	-0.0341	-0.1899
31	95	84	2	165	18	-31.5783	-2.3971	0.10927	0.0992	-0.7955	-0.273	0.1662	0.5392	-0.304	0.5339
32	88	80	2	190	60	2.0491	0.1567	0.00177	0.2587	0.0926	-0.0215	0.0121	-0.0078	-0.0167	0.0765
33	121	81	4	185	30	4.2222	0.286	0.00094	0.0528	0.0675	0.0079	-0.0199	-0.0385	0.0369	-0.0174
34	91	82	3	205	57	1.4146	0.1047	0.0006	0.2095	0.0539	-0.0171	0.0049	-0.006	0.0005	0.0398
35	93	79	3	140	35	-20.3213	-1.4488	0.03809	0.086	-0.4443	-0.0953	-0.1037	0.0475	0.2808	-0.2493

Figure 12 Summary of diagnostics for preliminary model

Figure 12 Shows the summary of diagnostics which was generated from SAS and excel was used to present the data in a well-organized manner. The h_{ii} are the leverage values. DFFITS and DFBETAS give the influence of a possible outlier. From *Figure 5*, the variance inflation values for all the four predictors are less than five. Hence, multicollinearity is not a serious issue.

Test	***Check column***	***Formula for cutoff value***	***Evaluation of the formula***	***Cutoff value***
Bonferroni y - outlier	Rstudent	$t_{(1-\frac{\alpha}{2n};n-p-1)}$	$t_{(1-\frac{0.1}{2*35};35-5-1)} = t_{(.9985714286;29)}$	3.25841
x - outlier	h_{ii}	$2p/n$	$2*5/35$	0.28571
Influence on $\hat{y}$	DFFITS	$2\sqrt{p/n}$	$2\sqrt{5/35}$	0.75593
Influence on all LSEs	CookDi	$F_{(0.5;P,n-p)}$	$F_{(0.5;5,35-5)} = F_{(0.5;5,30)}$	0.89019
Influence on individual LSE	DFBETAS	$2/\sqrt{n}$	$2/\sqrt{35}$	0.33806

Table 2 Cutoff values for outlier tests and influence for preliminary model

Based on the values shown in *Table 2*, "RStudent" and h_{ii} columns in *Figure 12* were verified for values exceeding the cutoff values. Only observation # 19 was found to be an *x* – outlier. There were no *y* – outliers.

The corresponding DFFITS, DFBETA and CookDi column values for observation #19 were verified to see if any of these were greater than their cutoff values shown in *Table 2*. It was found that this observation is influential only on the LSE of the predictor RPM (x_1). On further review, we find that the RPM value of 120 is the highest among the remaining RPM values. This was suspected to be an outlier during the review of scatterplots. However, since the DFBETA value is slightly above the cutoff value, this data point is being retained and we continue with further analysis.

Modified-Levene test

The dataset was split at a Predicted Heartrate value $\hat{y}$ = 127 because it splits the data set in half. As there is a lack of data points between heartrates of 90 – 110 and the points are clustered beyond 110, splitting it in half will give us a better idea of the variance.

group	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
1		17	8.7682	5.8107	1.4093	0	21.9457
2		18	7.1903	5.2423	1.2356	0.7934	18.7315
Diff (1-2)	Pooled		1.5779	5.5252	1.8686		
Diff (1-2)	Satterthwaite		1.5779		1.8743		

group	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
1		8.7682	5.7806	11.7558	5.8107	4.3276	8.8435
2		7.1903	4.5834	9.7973	5.2423	3.9338	7.8590
Diff (1-2)	Pooled	1.5779	-2.2239	5.3796	5.5252	4.4565	7.2727
Diff (1-2)	Satterthwaite	1.5779	-2.2391	5.3949			

Figure 13 SAS output showing the data split

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	33	0.84	0.4045
Satterthwaite	Unequal	32.165	0.84	0.4061

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	16	17	1.23	0.6769

Figure 14 SAS output for F-test and t-test

Below is the F-test to determine if equal or unequal variances need to be considered for the t-test. The significance level for this test is assumed to be 0.1.

$H_0 - \sigma_1 = \sigma_2$
$H_1 - \sigma_1 \neq \sigma_2$

P-value from *Figure 14* (Highlighted in red) = 0.6769

P-value > a (0.1)

Hence, we fail to reject H_0. The variances of the two groups are equal.

Now that we know about the variances to be considered for the t-test are equal, we conduct the t-test as follows.

H_0 – Means of d_{i1} and d_{i2} populations are equal
H_1 – Means are not equal

P-value from *Figure 14* (Highlighted in orange) = 0.4045

P-value > a (0.1)

Hence, we fail to reject H_0. This means the absolute deviations of the residuals around their group medians are equal (Weak conclusion).

This implies that the residuals have constant variance. This is in alignment with our analysis of residual vs. fitted value plot.

Test for Normality

Below is the SAS output for the correlation between residuals vs. z scores.

Pearson Correlation Coefficients, N = 35		
	e	enrm
e e(Y | x1,x2,x3,x4)	1.00000	0.99565
enrm Normal Scores	0.99565	1.00000

Figure 15 Correlation matrix of e vs. Z scores

H_0 – Normality is OK.
H_1 – Normality is violated.

$C_{(\propto=0.1;n=35)}$ = 0.974

From *Figure 15*, $\rho_{(e,z)}$= 0.99565

$\rho_{(e,z)} > C_{(\propto=0.1;n=35)}$ and hence, we fail to reject H_0.

Normality is OK (Weak conclusion).

Below is the preliminary model that satisfies the model assumptions:

$$\hat{y}_i = 134.428 + 0.196\, x_{i1} + 1.513\, x_{i2} - 0.066\, x_{i3} - 0.909\, x_{i4}$$

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	7298.42182	1824.60545	8.18	0.0001
Error	30	6693.74961	223.12499		
Corrected Total	34	13992			

Root MSE	14.93737	R-Square	0.5216
Dependent Mean	123.22857	Adj R-Sq	0.4578
Coeff Var	12.12168		

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Type I SS	Variance Inflation
Intercept	1	134.42796	18.55466	7.24	<.0001	531485	0
RPM	1	0.19601	0.21799	0.90	0.3757	70.26430	1.42226
Incline_level	1	1.51340	0.58235	2.60	0.0144	2495.00474	1.28566
Weights	1	-0.06654	0.09733	-0.68	0.4995	2084.41260	1.73582
Age	1	-0.90904	0.26384	-3.45	0.0017	2648.74018	1.47678

Figure 16 SAS output for the preliminary model

52.16% of the variability in Heartrate is explained by RPM of the equipment, Incline level of the equipment, Age and weight of the person using the equipment. The $adj\ R^2$ is less than R^2 which indicates that one or more predictors are not explaining much.

The regression is significant at 0.1 level because the p-value (highlighted in red) is less than α of 0.1. However, if we look at the p-values of the predictors in the parameter estimates table in *Figure 16*, RPM and Weight are not significant at 0.1 level. Hence, they could be removed one at a time in further analysis after interaction terms are explored and added to this preliminary model with four predictors.

III. Exploration of Interaction Terms

Since there are four predictors in the preliminary model, six interactions terms have been generated and analyzed to check if any of them can be added to the model. Partial regression plots involve regressing each interaction term against all the four predictors from the preliminary model. Then the residuals from the preliminary model are plotted against the residuals from the partial regression plots.

Figure 17 shows the partial regression plot for the interaction term x_1x_2. A downward trend can be seen in the plot. Hence, x_1x_2 can be added to the preliminary model. The partial regression plot for x_1x_3 in *Figure 18* does not show a clear trend. Hence, it can be ignored.

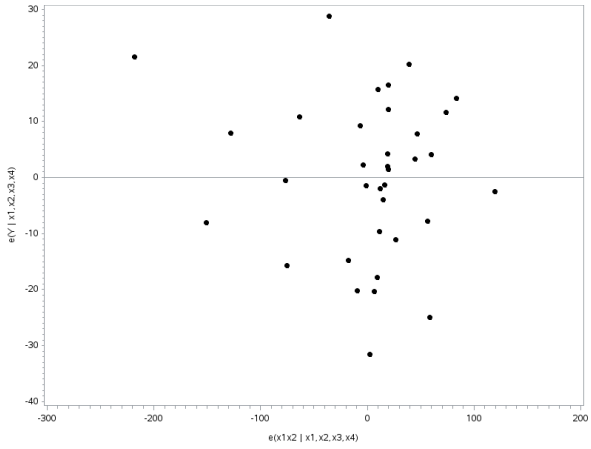

Figure 17 Partial regression plot for x_1x_2

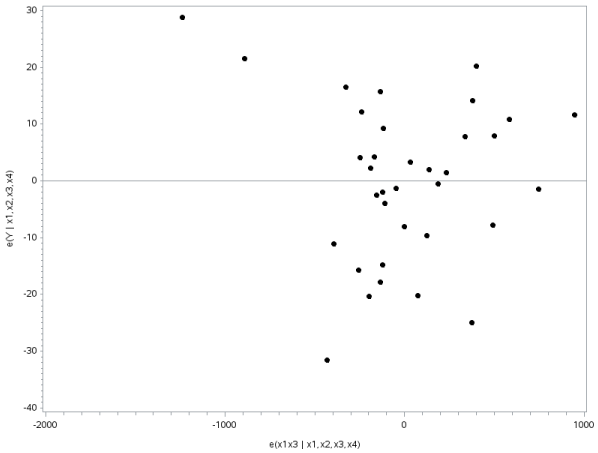

Figure 18 Partial regression plot for x_1x_3

Figure 19 shows the partial regression plot for x_1x_4. There is a possible downward trend in the plot. When it comes to x_2x_3, there is a possible upward trend as seen in *Figure 20*. Therefore, both x_1x_4 and x_2x_3 can possibly be added to the preliminary model.

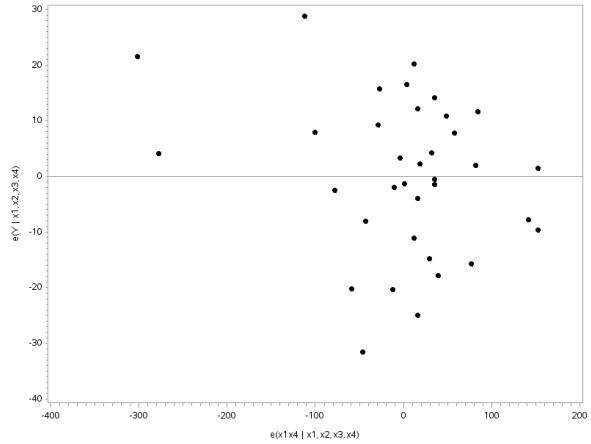

Figure 19 Partial regression plot for x_1x_4

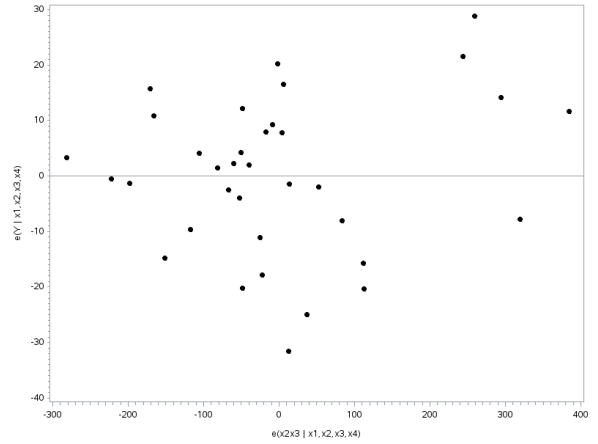

Figure 20 Partial regression plot for x_2x_3

Figure 21 and *Figure 22* are the partial regression plots for x_2x_4 and x_3x_4. Both of them do not show any clear trend. Hence, these interaction terms can be ignored.

In conclusion, based on the partial regression plots, x_1x_2, x_1x_4 and x_2x_3 are the three interaction terms that can be added to the preliminary model before trying to search for best models.

Prior to adding the interaction terms to the model, the predictors need to be standardized and then multiplied to generate standardized interaction terms.

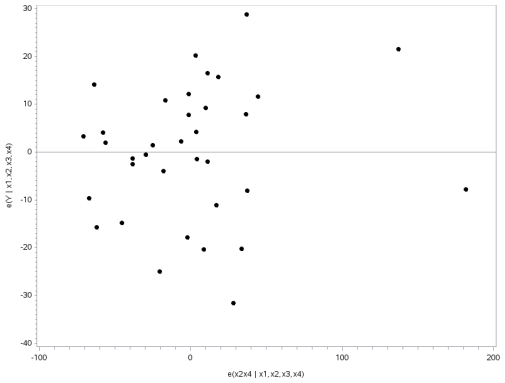

Figure 21 Partial regression plot for x_2x_4

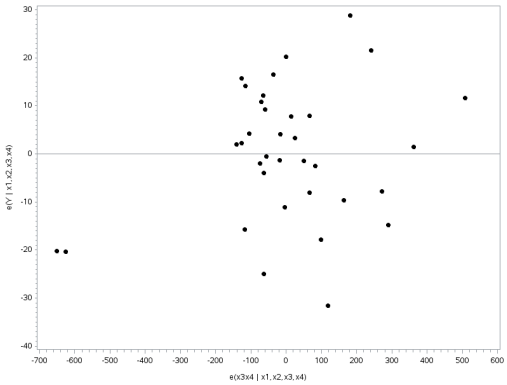

Figure 22 Partial regression plot for x_3x_4

Now that the standardized interaction terms have been generated, the correlation among the predictors and the interactions terms before and after standardizing will need to be checked to see if they have high multicollinearity.

Pearson Correlation Coefficients, N = 35											
	Heart_rate	RPM	Incline_level	Weights	Age	x1x2	x1x3	x1x4	x2x3	x2x4	x3x4
Heart_rate	1.00000	0.07086	0.39873	-0.21636	-0.64123	0.41043	-0.09912	-0.60776	0.33810	0.18574	-0.58627
RPM	0.07086	1.00000	-0.20532	0.33651	-0.09480	-0.03361	0.79679	0.28506	-0.14476	-0.24620	0.00661
Incline_level	0.39873	-0.20532	1.00000	0.18414	-0.14254	0.97070	-0.02442	-0.23151	0.96375	0.88309	-0.05668
Weights	-0.21636	0.33651	0.18414	1.00000	0.40988	0.21970	0.82770	0.49416	0.38539	0.32861	0.64295
Age	-0.64123	-0.09480	-0.14254	0.40988	1.00000	-0.17127	0.18562	0.92149	-0.04885	0.19186	0.95810
x1x2	0.41043	-0.03361	0.97070	0.21970	-0.17127	1.00000	0.10287	-0.18768	0.93809	0.82220	-0.07307
x1x3	-0.09912	0.79679	-0.02442	0.82770	0.18562	0.10287	1.00000	0.46703	0.14023	0.03672	0.39580
x1x4	-0.60776	0.28506	-0.23151	0.49416	0.92149	-0.18768	0.46703	1.00000	-0.12417	0.06630	0.91279
x2x3	0.33810	-0.14476	0.96375	0.38539	-0.04885	0.93809	0.14023	-0.12417	1.00000	0.91767	0.08511
x2x4	0.18574	-0.24620	0.88309	0.32861	0.19186	0.82220	0.03672	0.06630	0.91767	1.00000	0.27173
x3x4	-0.58627	0.00661	-0.05668	0.64295	0.95810	-0.07307	0.39580	0.91279	0.08511	0.27173	1.00000

Figure 23 Correlation matrix for response variable, predictors and interaction terms

As seen in *Figure 23,* the correlation of the three interaction terms with the predictors in the preliminary model have been highlighted in orange. Among these, two of the interaction terms, i.e., x_1x_2 and x_2x_3 are highly correlated with the predictor "Incline level". Also, x_1x_4 is highly correlated with the predictor "Age". The remaining correlation values are all low and are not a cause of concern.

If we look at the correlation of the standardized interaction terms with the predictors which are highlighted in green in in *Figure 24*, none of them are higher than 0.7. Hence, bringing in the interactions x_1x_2, x_1x_4 and x_2x_3 in standardized form is not going to cause any serious multicollinearity issues like variance inflation.

Pearson Correlation Coefficients, N = 35											
	Heart_rate	RPM	Incline_level	Weights	Age	stdx1x2	stdx1x3	stdx1x4	stdx2x3	stdx2x4	stdx3x4
Heart_rate	1.00000	0.07086	0.39873	-0.21636	-0.64123	-0.05298	0.04939	-0.01255	0.25086	0.37744	-0.32331
RPM	0.07086	1.00000	-0.20532	0.33651	-0.09480	-0.34367	-0.01587	-0.32519	-0.27110	-0.02578	-0.35348
Incline_level	0.39873	-0.20532	1.00000	0.18414	-0.14254	0.04290	-0.26375	-0.03942	0.29350	0.10386	-0.14868
Weights	-0.21636	0.33651	0.18414	1.00000	0.40988	-0.25402	-0.31913	-0.45228	-0.07790	-0.12439	-0.11778
Age	-0.64123	-0.09480	-0.14254	0.40988	1.00000	-0.02572	-0.30639	-0.26324	-0.13247	-0.47857	0.64257
stdx1x2	-0.05298	-0.34367	0.04290	-0.25402	-0.02572	1.00000	0.27858	0.46079	0.05621	-0.24759	0.11531
stdx1x3	0.04939	-0.01587	-0.26375	-0.31913	-0.30639	0.27858	1.00000	0.57548	-0.02225	0.06535	-0.05914
stdx1x4	-0.01255	-0.32519	-0.03942	-0.45228	-0.26324	0.46079	0.57548	1.00000	0.05659	-0.03444	0.07750
stdx2x3	0.25086	-0.27110	0.29350	-0.07790	-0.13247	0.05621	-0.02225	0.05659	1.00000	0.55128	0.10489
stdx2x4	0.37744	-0.02578	0.10386	-0.12439	-0.47857	-0.24759	0.06535	-0.03444	0.55128	1.00000	-0.21283
stdx3x4	-0.32331	-0.35348	-0.14868	-0.11778	0.64257	0.11531	-0.05914	0.07750	0.10489	-0.21283	1.00000

Figure 24 Correlation matrix for response variable, predictors and standardized interaction terms

IV. Model Search

Backwards Deletion

Backwards deletion was conducted by adding all the seven predictors i.e., x_1, x_2, x_3, x_4 and x_2x_3, x_1x_2, x_1x_4 in standardized form to the initial model. Then, predictor with the highest p-value was removed one-at a time as seen in *Figure 25*. An α of 0.1 was used to know if a p-value is high i.e., any p-value that is less than 0.1 cannot be removed in a step.

Summary of Backward Elimination							
Step	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	stdx1x2	6	0.0000	0.5597	6.0016	0.00	0.9683
2	RPM	5	0.0078	0.5520	4.4780	0.49	0.4880
3	stdx2x3	4	0.0067	0.5453	2.8872	0.43	0.5162
4	Weights	3	0.0110	0.5343	1.5603	0.72	0.4016
5	stdx1x4	2	0.0267	0.5076	1.1996	1.78	0.1919

Figure 25 Summary of backwards deletion output from SAS

Figure 27 shows the two final steps of the backwards deletion process. It can be seen in step 4 that x_1x_4 in standardized form was the last variable to be removed because its p-value was higher than 0.1. The final step gives a model with only "Age" and "Incline level" in the model. This model has been designated by letter "A".

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Type I SS	Variance Inflation
Intercept	1	143.77353	7.79691	18.44	<.0001	531485	0
Incline_level	1	1.27585	0.50973	2.50	0.0176	2224.50673	1.02074
Age	1	-1.02558	0.21548	-4.76	<.0001	4877.60668	1.02074

Figure 26 Regression output of Model A

As we can see in *Figure 26*, the p-value is less than 0.1 for both "Age" and "Incline level". The VIFs for these predictors are less than 5. Therefore, multicollinearity is nor a serious problem for "Model A".

Backward Elimination: Step 4

Variable Weights Removed: R-Square = 0.5343 and C(p) = 1.5603

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	7476.12246	2492.04082	11.86	<.0001
Error	31	6516.04897	210.19513		
Corrected Total	34	13992			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	145.93164	7.87172	72241	343.68	<.0001
Incline_level	1.22154	0.50527	1228.51235	5.84	0.0217
Age	-1.10580	0.22123	5251.47521	24.98	<.0001
stdx1x4	-5.15701	3.86606	374.00905	1.78	0.1919

Bounds on condition number: 1.1022, 9.6333

Backward Elimination: Step 5

Variable stdx1x4 Removed: R-Square = 0.5076 and C(p) = 1.1996

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	7102.11342	3551.05671	16.49	<.0001
Error	32	6890.05801	215.31431		
Corrected Total	34	13992			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	143.77353	7.79691	73213	340.03	<.0001
Incline_level	1.27585	0.50973	1348.94033	6.26	0.0176
Age	-1.02558	0.21548	4877.60668	22.65	<.0001

Bounds on condition number: 1.0207, 4.083

A

Figure 27 Final steps of backwards deletion

Best subsets

B

Number in Model	Adjusted R-Square	R-Square	C(p)	AIC	SBC	Variables in Model
1	0.3933	0.4112	5.1121	195.1450	198.25571	Age
1	0.1335	0.1590	20.5784	207.6217	210.73239	Incline_level

A

Number in Model	Adjusted R-Square	R-Square	C(p)	AIC	SBC	Variables in Model
2	0.4768	0.5076	1.1996	190.8870	195.55308	Incline_level Age
2	0.4119	0.4465	4.9449	194.9789	199.64498	Age stdx1x4

Number in Model	Adjusted R-Square	R-Square	C(p)	AIC	SBC	Variables in Model
3	0.4892	0.5343	1.5603	190.9336	197.15503	Incline_level Age stdx1x4
3	0.4676	0.5146	2.7686	192.3840	198.60535	Incline_level Age stdx2x3

X

Number in Model	Adjusted R-Square	R-Square	C(p)	AIC	SBC	Variables in Model
4	0.4847	0.5453	2.8872	192.0990	199.87573	Incline_level Weights Age stdx1x4
4	0.4817	0.5427	3.0449	192.2963	200.07308	Incline_level Age stdx1x4 stdx2x3

X

Number in Model	Adjusted R-Square	R-Square	C(p)	AIC	SBC	Variables in Model
5	0.4747	0.5520	4.4780	193.5816	202.91365	Incline_level Weights Age stdx1x4 stdx2x3
5	0.4723	0.5499	4.6051	193.7430	203.07514	RPM Incline_level Weights Age stdx1x4

X

Number in Model	Adjusted R-Square	R-Square	C(p)	AIC	SBC	Variables in Model
6	0.4654	0.5597	6.0016	194.9694	205.85688	RPM Incline_level Weights Age stdx1x4 stdx2x3
6	0.4564	0.5523	6.4551	195.5524	206.43980	Incline_level Weights Age stdx1x2 stdx1x4 stdx2x3

Figure 28 Best subsets output from SAS

In this process of model search, we began with only one predictor. The SAS procedure returned two best models with "Age" and "Incline level" in it. This process of finding two best models is continued with 2, 3, 4, 5 and 6 predictors in it.

If we observe the c_p values, the least value is found in the models with three predictors. However, the AIC and SBC is the lowest for the model with two predictors "Age" and "Incline level" in it. This model is similar to the "Model A" that was seen as the result of the "*Backwards deletion*" process. If we observe the $adj\ R^2$ value, it starts decreasing after 4 predictors are added. Hence, the models with four predictors and beyond cannot be considered for model search and they have been marked with letter "X".

The three-predictor model with Incline level, Age and x_1x_4 in standardized form was rejected in "*Backwards deletion*" because standardized x_1x_4 was not significant. The next three-predictor model with Incline level, Age and x_2x_3 in standardized form is not acceptable because the standardized x_2x_3 term is insignificant as seen in *Figure 29*.

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Type I SS	Variance Inflation
Intercept	1	143.71258	7.86546	18.27	<.0001	531485	0
Incline_level	1	1.17541	0.53559	2.19	0.0358	2224.50673	1.10752
Age	1	-1.01156	0.21836	-4.63	<.0001	4877.60668	1.03019
stdx2x3	1	1.86060	2.77731	0.67	0.5079	98.32748	1.10439

Figure 29 Regression output for three-predictor model with Age, Incline level and x_2x_3

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Type I SS	Variance Inflation
Intercept	1	156.39652	7.05483	22.17	<.0001	531485	0
Age	1	-1.19064	0.23438	-5.08	<.0001	5753.17308	1.07446
stdx1x4	1	-5.91014	4.13491	-1.43	0.1626	494.43704	1.07446

Figure 30 Regression output for a two-predictor model with Age and x_1x_4

The model with two predictors, Age and standardized x_1x_4 is also not acceptable because the term x_1x_4 is insignificant as seen in *Figure 30*. If we look at the remaining options for choosing a second-best model, it has to be "Model B" because in comparison to the model with just the Incline level in it, Model B has a better R^2 and there will not be multicollinearity in this model because it only has one predictor. It can be seen in *Figure 31* that "Age" is significant at 0.1 level.

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Type I SS	Variance Inflation
Intercept	1	154.44378	7.02981	21.97	<.0001	531485	0
Age	1	-1.10245	0.22966	-4.80	<.0001	5753.17308	1.00000

Figure 31 Regression output for Model B

Stepwise Regression

The stepwise regression was conducted by considering possible addition of seven predictors i.e., x_1, x_2, x_3, x_4 and x_2x_3, x_1x_2, x_1x_4 in standardized form. An α of 0.1 was used in the process. As seen in *Figure 32*, the first variable which was added to the model with no predictors was Age (x_4). It is similar to "Model B" that was seen in the "*Best subsets"* procedure.

Stepwise Selection: Step 1

Variable Age Entered: R-Square = 0.4112 and C(p) = 5.1121

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	5753.17308	5753.17308	23.04	<.0001
Error	33	8238.99835	249.66662		
Corrected Total	34	13992			

B

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	154.44378	7.02981	120507	482.67	<.0001
Age	-1.10245	0.22966	5753.17308	23.04	<.0001

Figure 32 Step 1 of the Stepwise Regression output from SAS

As seen in *Figure 33,* in the second step, Incline level (x_2) was added. At this point, no more predictors could be added because they were insignificant when added. The final model suggested by stepwise regression is "Model A" which is similar to the result from "*Backwards deletion"* and *"Best subset"*.

Stepwise Selection: Step 2

Variable Incline_level Entered: R-Square = 0.5076 and C(p) = 1.1996

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	7102.11342	3551.05671	16.49	<.0001
Error	32	6890.05801	215.31431		
Corrected Total	34	13992			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	143.77353	7.79691	73213	340.03	<.0001
Incline_level	1.27585	0.50973	1348.94033	6.26	0.0176
Age	-1.02558	0.21548	4877.60668	22.65	<.0001

Bounds on condition number: 1.0207, 4.083

All variables left in the model are significant at the 0.1000 level.

No other variable met the 0.1000 significance level for entry into the model.

A

Summary of Stepwise Selection								
Step	Variable Entered	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	Age		1	0.4112	0.4112	5.1121	23.04	<.0001
2	Incline_level		2	0.0964	0.5076	1.1996	6.26	0.0176

Figure 33 Final step and summary of Stepwise Regression output from SAS

The best models chosen from model search are as follows:

Model A

$$\hat{y}_i = 143.773 + 1.276\, x_{i2} - 1.025\, x_{i4}$$

Model B

$$\hat{y}_i = 154.444 - 1.102\, x_{i4}$$

V. Model Selection

Model A – Model Assumption Verification

Model Form

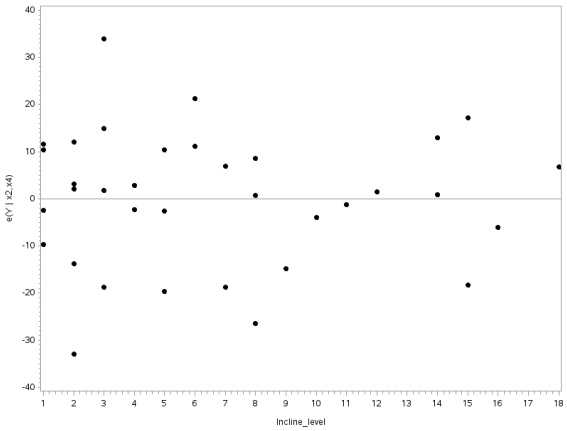

Figure 34 Residual vs. Incline level

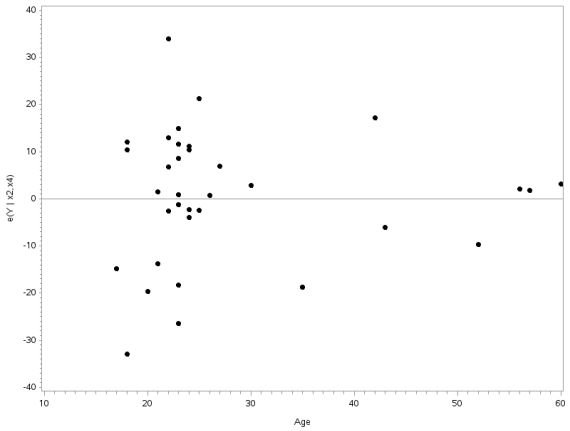

Figure 35 Residual vs. Age

As seen in *Figure 35,* there are not many data points 30 and 50 age groups. Most of the data points are clustered between the age groups of 20 and 30. Also, there could be outliers between the age groups of 50 and 60 as noticed previously in the preliminary model form verification. None of the residual plots shown in *Figure 34* and *Figure 35* show curvature. Hence, MLR model form is adequate.

Constant Variance

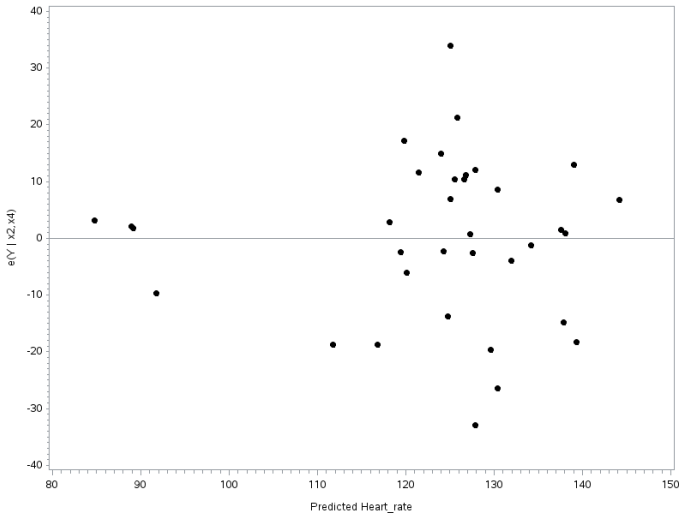

Figure 36 Residual vs. Predicted Heartrate

Figure 36 shows the "*e* vs. $\hat{y}$" plot. It can be seen that the points are randomly scattered. The points are clustered beyond heartrate of 110 beats per minute. There is no funnel and hence, variance appears to be

constant. If there were more data points between the heartrates of 90 and 110, we might see a funnel shape. Curvature is also not seen in this plot. No observations appear to be an outlier.

Normality

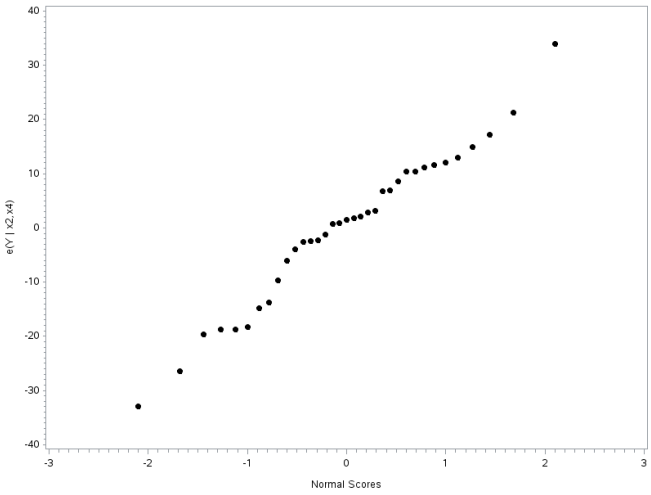

Figure 37 Normal Probability Plot

The normal probability plot in *Figure 37* appears to be slightly right skewed. However, overall, it is quite straight and close to normality.

Model A – Diagnostic Checks

Observation #	Response	Predictors		Residual	RStudent	cookdi	hii	DFFITS	DFBETAS		
	Heart_rate	Incline_level	Age						Intercept	Incline_level	Age
1	139	8	23	8.6079	0.5913	0.00442	0.0358	0.1139	0.0567	0.0214	-0.043
2	125	5	22	-2.5901	-0.1776	0.00048	0.042	-0.0372	-0.0297	0.0128	0.0183
3	139	3	23	14.9872	1.0516	0.02082	0.0536	0.2503	0.2045	-0.1494	-0.1039
4	111	2	21	-13.7881	-0.9743	0.0243	0.0712	-0.2698	-0.2355	0.179	0.1319
5	98	7	35	-18.8093	-1.3226	0.02293	0.0387	-0.2653	0.0481	-0.0349	-0.1347
6	133	1	23	11.5389	0.8145	0.01881	0.0777	0.2363	0.1926	-0.1761	-0.0899
7	137	1	18	10.411	0.742	0.02037	0.0986	0.2454	0.226	-0.1705	-0.1403
8	152	14	22	12.9273	0.9242	0.0302	0.0955	0.3003	-0.0072	0.235	-0.0548
9	138	6	24	11.1852	0.7704	0.00695	0.0335	0.1435	0.0939	-0.025	-0.0522
10	159	3	22	33.9616	2.5861	0.1137	0.0567	0.6338	0.5366	-0.3735	-0.2951
11	122	4	24	-2.2631	-0.1552	0.00037	0.0428	-0.0328	-0.025	0.0161	0.0121
12	139	12	21	1.4534	0.101	0.00026	0.0689	0.0275	0.0045	0.0178	-0.0085
13	137	15	42	17.163	1.2971	0.11212	0.1696	0.5861	-0.3906	0.4533	0.3449
14	128	8	26	0.6847	0.0467	0.00002	0.0315	0.0084	0.0027	0.002	-0.0013
15	147	6	25	21.2108	1.4973	0.02361	0.0318	0.2713	0.1634	-0.0453	-0.079
16	121	15	23	-18.323	-1.3417	0.07399	0.1122	-0.477	0.0552	-0.3969	0.0523
17	114	16	43	-6.0873	-0.4582	0.01798	0.2004	-0.2293	0.1567	-0.182	-0.1342
18	140	2	18	12.1351	0.8609	0.02304	0.0847	0.2618	0.2431	-0.165	-0.157
19	117	1	25	-2.41	-0.1679	0.00076	0.0723	-0.0469	-0.0353	0.0355	0.0134
20	136	5	24	10.4611	0.721	0.00675	0.0369	0.1412	0.1019	-0.0489	-0.0526
21	139	14	23	0.9529	0.0672	0.00016	0.094	0.0216	-0.0014	0.0172	-0.0029
22	123	9	17	-14.8214	-1.0426	0.02264	0.0589	-0.2609	-0.1608	-0.062	0.1661
23	132	7	27	6.9861	0.4773	0.00233	0.029	0.0825	0.0308	0.0044	-0.0085
24	133	11	23	-1.2196	-0.0841	0.00014	0.0541	-0.0201	-0.0032	-0.0121	0.0049
25	104	8	23	-26.3921	-1.9055	0.04152	0.0358	-0.3672	-0.1827	-0.0689	0.1385
26	128	10	24	-3.9182	-0.2691	0.00114	0.044	-0.0577	-0.0121	-0.0295	0.0129
27	91	2	56	2.107	0.1582	0.00217	0.2013	0.0794	-0.0356	-0.0184	0.0679
28	110	5	20	-19.6412	-1.3924	0.03221	0.0488	-0.3154	-0.2689	0.107	0.186
29	151	18	22	6.8239	0.5082	0.01961	0.182	0.2397	-0.0468	0.214	-0.0205
30	82	1	52	-9.7194	-0.721	0.0357	0.1687	-0.3248	0.1048	0.1162	-0.2529
31	95	2	18	-32.8649	-2.5311	0.16898	0.0847	-0.7698	-0.7147	0.4851	0.4617
32	88	2	60	3.2093	0.2488	0.00708	0.2498	0.1436	-0.072	-0.0274	0.127
33	121	4	30	2.8904	0.1977	0.00052	0.0371	0.0388	0.0168	-0.0179	0.0023
34	91	3	57	1.8568	0.1399	0.00176	0.2069	0.0715	-0.0365	-0.0105	0.0633
35	93	3	35	-18.7059	-1.3238	0.03054	0.0508	-0.3062	-0.0548	0.1536	-0.1088

Figure 38 Summary of diagnostics for Model A

Figure 38 shows the summary of diagnostics which was generated from SAS and excel was used to present the data in a well-organized manner. The h_{ii} are the leverage values. DFFITS and DFBETAS give the influence of a possible outlier.

Test	***Check column***	***Formula for cutoff value***	***Evaluation of the formula***	***Cutoff value***
Bonferroni y - outlier	Rstudent	$t_{(1-\frac{\alpha}{2n};n-p-1)}$	$t_{(1-\frac{0.1}{2*35};35-3-1)} = t_{(.9985714286;31)}$	3.23927
x - outlier	h_{ii}	$2p/n$	$2*3/35$	0.17143
Influence on $\hat{y}$	DFFITS	$2\sqrt{p/n}$	$2\sqrt{3/35}$	0.58554
Influence on all LSEs	CookDi	$F_{(0.5;P,n-p)}$	$F_{(0.5;3,35-3)} = F_{(0.5;3,32)}$	0.80573
Influence on individual LSE	DFBETAS	$2/\sqrt{n}$	$2/\sqrt{35}$	0.33806

Table 3 Cutoff values for outlier tests and influence for Model A

Based on the values shown in *Table 3*, "RStudent" and h_{ii} columns in *Figure 38* were verified for values exceeding the cutoff values. Observation # 17 and 29 were found to be *x* – outliers probably because of the high Incline levels of 16 and 18. As suspected earlier, observation # 27, 32 and 34 were also found to be *x* – outliers because the age group is between 50 and 60. There were no *y* – outliers.

The corresponding DFFITS, DFBETA and CookDi column values for these outliers were verified to see if any of these were greater than their cutoff values shown in *Table 3*. It was found that none of them are influential. Hence, these outliers are not a cause of concern and the data can be retained as is.

Model B – Model Assumption Verification

Model Form and Constant Variance

As seen in *Figure 39* and *Figure 40*, none of them show curvature. The points in both the plots appear to be randomly scattered. Hence a linear model form is adequate.

The points are clustered between age groups of 20 and 30 in *Figure 40* and beyond a Heartrate of 110 in *Figure 39*. However, there is no clear funnel shape as there is a point with a residual of 30 at a Heartrate of 105. Hence the variance might be constant and having more data points between 90 and 110 will help assessing the residual plots in a better way.

As with the previous model, there could be few outliers between the age group of 50 and 60. However, it can only be confirmed through diagnostic checks which will be discussed further.

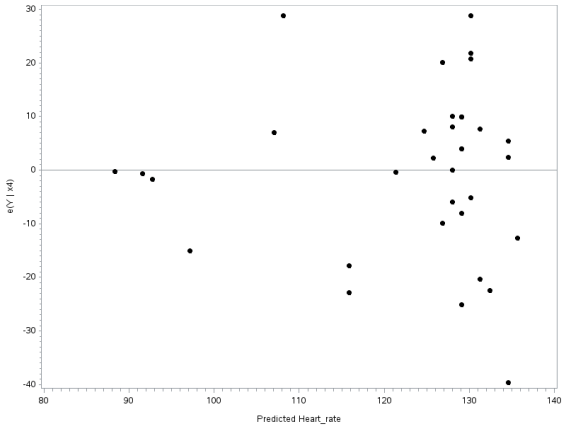

Figure 39 Residual vs. Predicted Heartrate

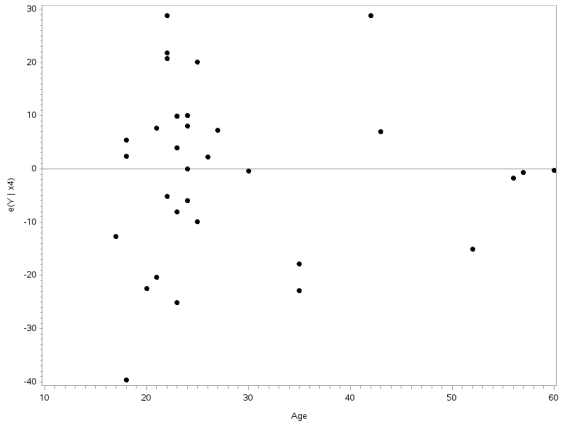

Figure 40 Residual vs. Age

Normality

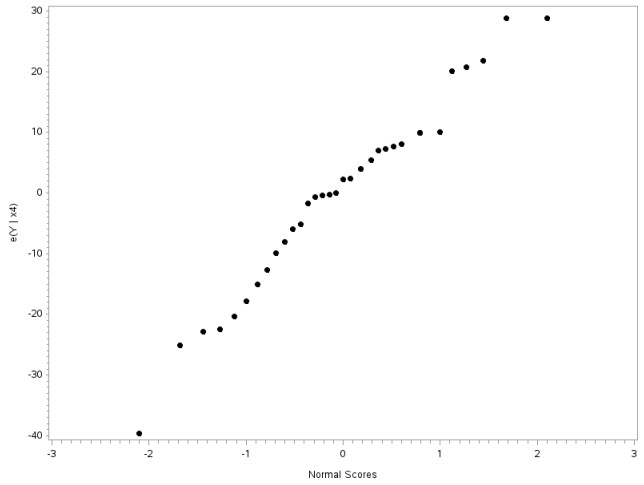

Figure 41 Normal Probability Plot

The normal probability plot in *Figure 41* appears to be slightly left skewed. However, overall, it is quite straight and close to normality.

Model B – Diagnostic Checks

Figure 42 shows the summary of diagnostics which was generated from SAS and excel was used to present the data in a well-organized manner. The h_{ii} are the leverage values. DFFITS and DFBETAS give the influence of a possible outlier.

Based on the values shown in *Table 4*, "RStudent" and h_{ii} columns in *Figure 42* were verified for values exceeding the cutoff values. As suspected earlier, observation # 27, 30, 32 and 34 were also found to be *x* – outliers because the age group is between 50 and 60. There were no *y* – outliers.

Observation #	Response Heart_rate	Predictor Age	Residual	RStudent	cookdi	hii	DFFITS	DFBETAS Intercept	DFBETAS Age
1	139	23	9.9127	0.6326	0.00729	0.0345	0.1197	0.0874	-0.0497
2	125	22	-5.1898	-0.3301	0.00215	0.037	-0.0647	-0.0502	0.0309
3	139	23	9.9127	0.6326	0.00729	0.0345	0.1197	0.0874	-0.0497
4	111	21	-20.2922	-1.3256	0.03567	0.0399	-0.2701	-0.2199	0.1438
5	98	35	-17.8579	-1.1582	0.02623	0.038	-0.2302	0.0303	-0.1148
6	133	23	3.9127	0.2484	0.00114	0.0345	0.047	0.0343	-0.0195
7	137	18	2.4004	0.1536	0.00065	0.051	0.0356	0.032	-0.0236
8	152	22	21.8102	1.4286	0.038	0.037	0.28	0.2171	-0.1336
9	138	24	10.0151	0.6386	0.00698	0.0325	0.117	0.0794	-0.0407
10	159	22	28.8102	1.9336	0.06631	0.037	0.379	0.2938	-0.1808
11	122	24	-5.9849	-0.3801	0.00249	0.0325	-0.0697	-0.0472	0.0242
12	139	21	7.7078	0.4921	0.00515	0.0399	0.1003	0.0816	-0.0534
13	137	42	28.8593	1.9733	0.13088	0.0681	0.5336	-0.2449	0.4066
14	128	26	2.22	0.1405	0.00031	0.0297	0.0246	0.0136	-0.0048
15	147	25	20.1176	1.3071	0.02666	0.0309	0.2334	0.1444	-0.064
16	121	23	-8.0873	-0.5151	0.00485	0.0345	-0.0974	-0.0711	0.0405
17	114	43	6.9618	0.4523	0.00839	0.0741	0.128	-0.0626	0.1003
18	140	18	5.4004	0.3461	0.00331	0.051	0.0803	0.0721	-0.0533
19	117	25	-9.8824	-0.6295	0.00643	0.0309	-0.1124	-0.0696	0.0308
20	136	24	8.0151	0.5099	0.00447	0.0325	0.0935	0.0634	-0.0325
21	139	23	9.9127	0.6326	0.00729	0.0345	0.1197	0.0874	-0.0497
22	123	17	-12.7021	-0.8232	0.02015	0.0556	-0.1998	-0.1833	0.1393
23	132	27	7.3225	0.4647	0.0033	0.0289	0.0802	0.0386	-0.009
24	133	23	3.9127	0.2484	0.00114	0.0345	0.047	0.0343	-0.0195
25	104	23	-25.0873	-1.6581	0.0467	0.0345	-0.3136	-0.2289	0.1303
26	128	24	0.0151	0.00096	0	0.0325	0.0002	0.0001	-0.0001
27	91	56	-1.7063	-0.1182	0.0017	0.1905	-0.0573	0.0405	-0.0529
28	110	20	-22.3947	-1.4745	0.04737	0.0432	-0.3132	-0.2653	0.1822
29	151	22	20.8102	1.3592	0.0346	0.037	0.2664	0.2065	-0.1271
30	82	52	-15.1162	-1.0371	0.09253	0.1471	-0.4307	0.2855	-0.3866
31	95	18	-39.5996	-2.8334	0.17802	0.051	-0.6572	-0.5901	0.4361
32	88	60	-0.2965	-0.0212	0.00007	0.2407	-0.0119	0.0088	-0.0112
33	121	30	-0.3701	-0.0234	0.00001	0.0292	-0.0041	-0.001	-0.0006
34	91	57	-0.6039	-0.0421	0.00023	0.2024	-0.0212	0.0152	-0.0197
35	93	35	-22.8579	-1.5028	0.04298	0.038	-0.2987	0.0393	-0.1489

Figure 42 Summary of Diagnostics for Model B

The corresponding DFFITS, DFBETA and CookDi column values for these outliers were verified to see if any of these were greater than their cutoff values shown in *Table 4*. Only Observation # 30 was found to be influential on the LSE of Age. However, the influence value is slightly higher than the cutoff value. It might be acceptable to retain the point.

Test	***Check column***	***Formula for cutoff value***	***Evaluation of the formula***	***Cutoff value***
y - outlier	Rstudent	$t_{(1-\frac{\alpha}{2n};n-p-1)}$	$t_{(1-\frac{0.1}{2*35};35-2-1)} = t_{(.9985714286;32)}$	3.23066
x - outlier	h_{ii}	$2p/n$	$2 * 2/35$	0.11428
Influence on $\hat{y}$	DFFITS	$2\sqrt{p/n}$	$2\sqrt{2/35}$	0.47809
Influence on all LSEs	CookDi	$F_{(0.5;P,n-p)}$	$F_{(0.5;2,35-2)} = F_{(0.5;2,33)}$	0.70791
Influence on individual LSE	DFBETAS	$2/\sqrt{n}$	$2/\sqrt{35}$	0.33806

Table 4 Cutoff values for outlier tests and influence - Model B

	Model A	***Model B***
# Of predictors	2	1
R^2	0.5076	0.4112
$adj\ R^2$	0.4768	0.3933
***e* vs. $\hat{y}$**	Constant variance	Constant variance
***e* vs. Predictors**	No curvature seen	No curvature seen
NPP	Slightly right skewed. Close to normality.	Slightly left skewed. Close to normality.
VIF	All VIFs are less than 5.	Only one predictor in model. VIF not an issue.
c_p	1.996	5.112
AIC	190.887	195.145
SBC	195.553	198.255
***x* - outlier**	5	4
***y* - outlier**	none	none
Influence	None of the outliers are influential.	One outlier influential.

Table 5 Comparison of Model A and Model B

Based on the comparison of the models, ***Model A is the best model***. This is because it has higher R^2 and $adjR^2$. Also, the c_p, AIC and SBC for Model A are lower when compared to Model B. Apart from this, though Model A has outliers, none of them are influential.

VI. Final Multiple Linear Regression Model

The final model that is selected is as follows:

$$\hat{y}_i = 143.773 + 1.276\, x_{i2} - 1.025\, x_{i4}$$

The parameter b_2 (1.276) indicates the change in the mean *Heartrate* per unit increase in the *Incline level* when *Age* is held constant.

The parameter b_4 (-1.025) indicates the change in the mean *Heartrate* per unit increase in the *Age* when *Incline level* is held constant.

The REG Procedure
Model: MODEL1
Dependent Variable: Heart_rate

Number of Observations Read	36
Number of Observations Used	35
Number of Observations with Missing Values	1

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	7102.11342	3551.05671	16.49	<.0001
Error	32	6890.05801	215.31431		
Corrected Total	34	13992			

Root MSE	14.67359	R-Square	0.5076
Dependent Mean	123.22857	Adj R-Sq	0.4768
Coeff Var	11.90762		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	143.77353	7.79691	18.44	<.0001
Incline_level	1	1.27585	0.50973	2.50	0.0176
Age	1	-1.02558	0.21548	-4.76	<.0001

Figure 43 ANOVA of Model A

50.76% of the variability in Heartrate is explained by the Incline level of the equipment and the Age of the person using the equipment. The $adj\ R^2$ is less than R^2 which indicates that one of the predictors is not explaining much.

The regression is significant at 0.1 level because the p-value is less than 0.1 as seen in the ANOVA section of the *Figure 43*. When it comes to parameter estimates, both Age and Incline level are significant at 0.1 level. If we were to consider an α of 0.01, Incline level becomes insignificant. This makes sense because, Age is known to affect a person's heartrate during physical activity. Hence, most people using exercise equipment are recommended to follow the instructions on the equipment with regards to heartrate based on their Age to ensure the person does not strain his heart. However, Incline level is also going to affect the Heartrate because, as the Incline level increases, the difficulty level increases thereby requiring more effort resulting in an increase in the heartrate.

Joint CI for parameters

An α of 0.1 has been used in the calculations.

We are 90% confident that the β_2 is in the interval (0.2367,2.3132) and β_4 is in the interval (-1.4645, -0.5866) simultaneously.

Since there was no data for Heartrate of a person aged 30 and using the equipment at an incline level of 13, we chose these values for estimating CI, CB and PI. Hence the $x_h$vector is (1,13,30). These interval values have been calculated at an α of 0.05.

CI at x_h

We are 95% confident that the mean Heartrate when a person aged 30 years is using an equipment at an Incline level of 13 is between 115.353 and 133.575 beats per minute.

CB at x_h

The confidence band values when a person aged 30 years is using an equipment at an Incline level of 13 is (111.488,137.44).

PI at x_h

We predict with 95% confidence that the actual Heartrate when a person aged 30 years is using an equipment at an Incline level of 13 is between 93.217 and 155.711 beats per minute.

VII. Final Discussion

Model A with two predictors, Age of a person and Incline level of the equipment being used by the person is the best model based on our analysis. The predictors in this model explain only 50.76% of the variability in Heartrate. This model had 5 outliers, however, none of them were influential and hence they have been retained. These five data points are probably being flagged as outliers because of lack of data points. Hence, more data points could help improve the model. Apart from this, we can consider adding other predictors like the BMI (Body Mass Index) as it helps classify if a person is obese because obesity also could affect the heartrate of a person during workout. Smoking habits can also be considered in a future analysis as a binary predictor.