Title: Oxygen Consumption of High-Performance Athletes.

Data Summary

The fitness dataset consisted of 31 observations and 7 variables. Individuals were given their age and weight with measurement of their maximum pulse rate, rest pulse rate, run pulse rate, run time and oxygen consumption after running.

The Summary of data:

```
      AGE              WEIGHT          OXYGEN          RUNTIME          RATEREST    
 Min.   :38.00   Min.   :59.08   Min.   :37.39   Min.   : 8.17   Min.   :40.00  
 1st Qu.:44.00   1st Qu.:73.20   1st Qu.:44.97   1st Qu.: 9.78   1st Qu.:48.00  
 Median :48.00   Median :77.45   Median :46.77   Median :10.47   Median :52.00  
 Mean   :47.68   Mean   :77.44   Mean   :47.38   Mean   :10.59   Mean   :53.45  
 3rd Qu.:51.00   3rd Qu.:82.33   3rd Qu.:50.13   3rd Qu.:11.27   3rd Qu.:58.50  
 Max.   :57.00   Max.   :91.63   Max.   :60.06   Max.   :14.03   Max.   :70.00  
    RATERUN         RATEMAX     
 Min.   :146.0   Min.   :155.0  
 1st Qu.:163.0   1st Qu.:168.0  
 Median :170.0   Median :172.0  
 Mean   :169.6   Mean   :173.8  
 3rd Qu.:176.0   3rd Qu.:180.0  
 Max.   :186.0   Max.   :192.0  
```

The Correlation between variables:

	AGE	WEIGHT	OXYGEN	RUNTIME	RATEREST	RATERUN	RATEMAX
AGE	1.00	-0.23	-0.30	0.19	-0.16	-0.34	-0.43
WEIGHT	-0.23	1.00	-0.16	0.14	0.04	0.18	0.25
OXYGEN	-0.30	-0.16	1.00	-0.86	-0.40	-0.40	-0.24
RUNTIME	0.19	0.14	-0.86	1.00	0.45	0.31	0.23
RATEREST	-0.16	0.04	-0.40	0.45	1.00	0.35	0.31
RATERUN	-0.34	0.18	-0.40	0.31	0.35	1.00	0.93
RATEMAX	-0.43	0.25	-0.24	0.23	0.31	0.93	1.00

Statement of Problem

Form the summary of the data, we can see that the median and mean values of each variable are relative close; therefore, we can conclude that there is no unusual observations in the data. So, we want to exam the underlying relationship between variables using principal component analysis.

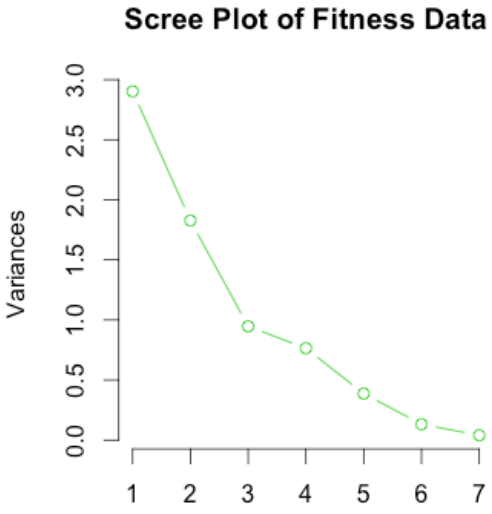
From the correlation test, oxygen is negatively correlated with runtime, rate at rest and rate after running. It seen that oxygen consumption after running is depended on some variables. Therefore, we want to find the significant predictors of oxygen consumption after running.

Analysis

Principal Component Analysis without RATEMAX since it is highly correlated with RATERUN.

```
pca = prcomp(fitness[,-7], scale =T)
```

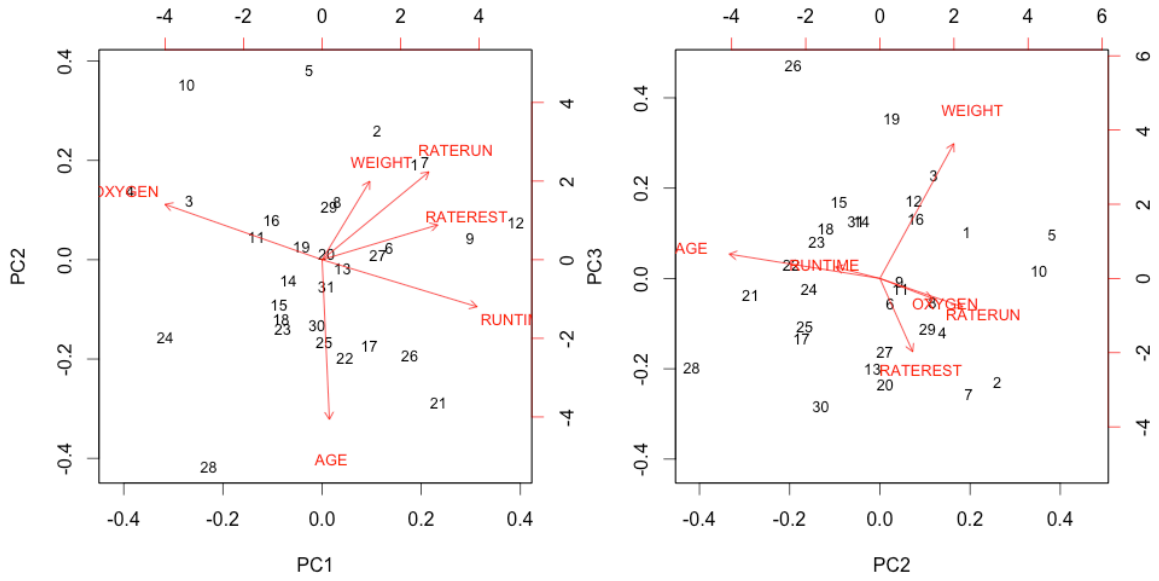
```
> summary(pca)
Importance of components:
                          PC1    PC2    PC3    PC4     PC5     PC6
Standard deviation     1.5746 1.2237 0.9653 0.7861 0.60607 0.32603
Proportion of Variance 0.4132 0.2495 0.1553 0.1030 0.06122 0.01772
Cumulative Proportion  0.4132 0.6628 0.8181 0.9211 0.98228 1.00000
```



Based on the scree plot and the cumulative proportion of variance, the first three principal components are sufficient for this data, which explains 0.811 of variance.

Coefficient of first three components:

	PC1	PC2	PC3
AGE	0.03	-0.75	0.15
WEIGHT	0.17	0.37	0.84
OXYGEN	-0.57	0.26	-0.12
RUNTIME	0.56	-0.22	0.07
RATEREST	0.42	0.16	-0.46
RATERUN	0.39	0.41	-0.18



The first principal component explained 0.431 of the variance in the data. It is mainly measurement of oxygen consumption after running and amount of time ran. As runtime increase, oxygen consumption tends to decrease after running.

The second component explained about 0.249 of the variance. It measured the age and run pulse rate. As age increases, run pulse rate tends to decrease.

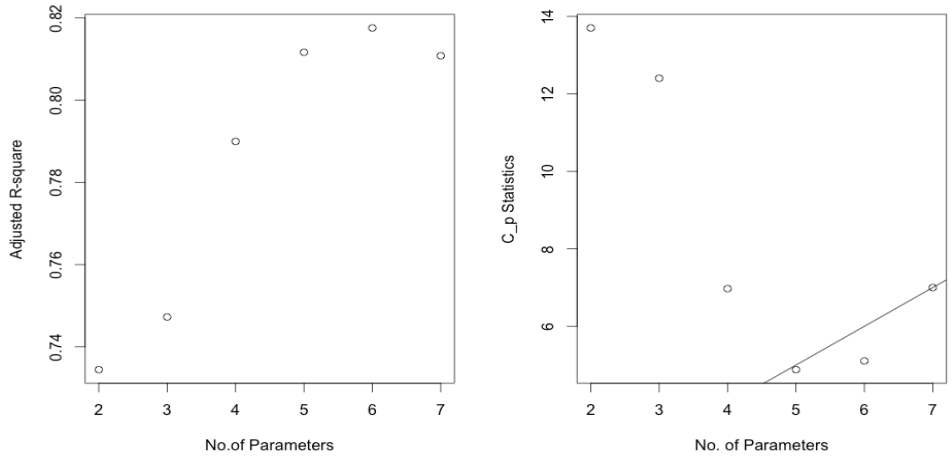
The third component explained 0.1553 of the variance. It is significantly measure the weight that is moderately correlated to rate at rest. As weight increase the rate at rest tends to be slower.

Variable Selection

In this analysis, we want find significant predictors of oxygen consumption using Mallow's C_p statistic. For Mallow's statistic, we want to find the number of variables to include in the model based on the fit of the model (R^2).

```
> all<- regsubsets(OXYGEN ~., data=fitness)
> (rs<-summary(all))
Subset selection object
Call: regsubsets.formula(OXYGEN ~ ., data = fitness)
6 Variables  (and intercept)
Selection Algorithm: exhaustive
         AGE WEIGHT RUNTIME RATEREST RATERUN RATEMAX
1  ( 1 ) " " " "    "*"     " "      " "     " "
2  ( 1 ) "*" " "    "*"     " "      " "     " "
3  ( 1 ) "*" " "    "*"     " "      "*"     " "
4  ( 1 ) "*" " "    "*"     " "      "*"     "*"
5  ( 1 ) "*" "*"    "*"     " "      "*"     "*"
6  ( 1 ) "*" "*"    "*"     "*"      "*"     "*"

> plot(2:7,rs$adjr2, xlab="No.of Parameters", ylab="Adjusted R-square")
> plot(2:7,rs$cp, xlab="No. of Parameters", ylab="C_p Statistics")
> abline(0,1)
```



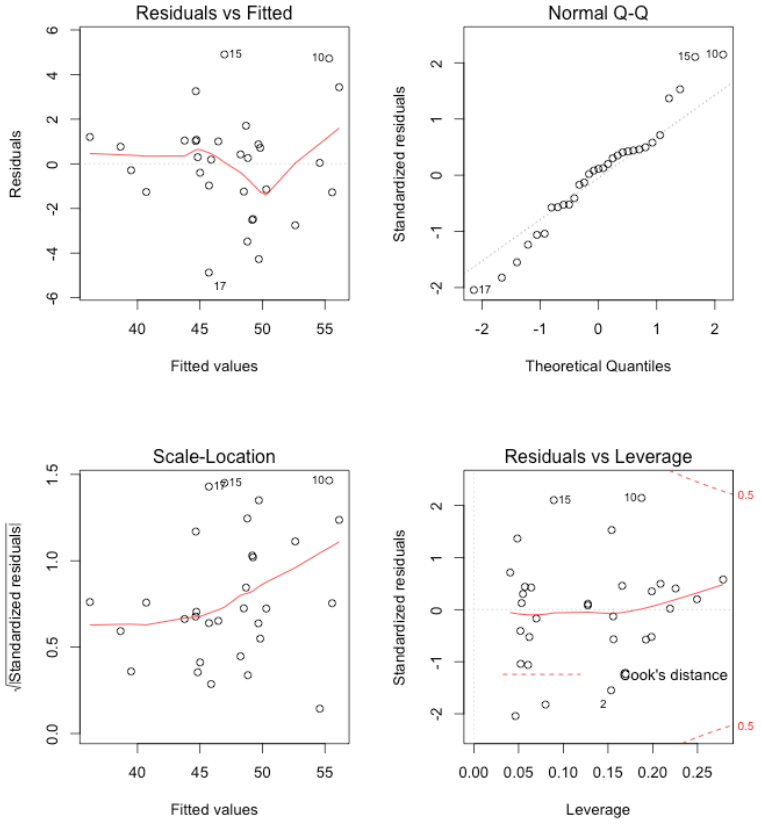
Based on the Mallow C_p Statistics, the number of parameters to be included in the model is five with adjusted R^2 approximate 81. These variables are AGE, RUNTIME, RATERUN, and RATEMAX. Then, we fit the model to further investigate their significance.

```
> mod = lm(OXYGEN ~AGE +RUNTIME + +RATERUN +RATEMAX, fitness)
> summary(mod)

Call:
lm(formula = OXYGEN ~ AGE + RUNTIME + +RATERUN + RATEMAX, data =
fitness)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 98.12654   11.78929   8.323 8.35e-09 ***
AGE         -0.19744    0.09567  -2.064   0.0491 *
RUNTIME     -2.76787    0.34064  -8.125 1.32e-08 ***
RATERUN     -0.34852    0.11754  -2.965   0.0064 **
RATEMAX      0.27098    0.13366   2.027   0.0530 .

Residual standard error: 2.312 on 26 degrees of freedom
Multiple R-squared:  0.8368,  Adjusted R-squared:  0.8117
```

Model Diagnostic



Based on the Residuals vs. Fitted plot we can tell variances are constant. From the q-q plot, we can tell that errors are normally distributed. And from Residual vs. Leverage plot, we can't detect ant outliners. Therefore, this model is fitted well.

From the summary of the model, we can see that the model with AGE,, RUNTIME, RATERUN, and RATEMAX is fit with high-adjusted R^2 (0.811) and all variables are significant level less than 0.05 except for RATEMAX. Since, RATEMAX and RATERUN are highly correlated (0.93), they are believed to have similar affect to the model, this is also confirmed with t test with p-value = 0.10 . Therefore, we remove RATEMAX from the model.

```
> t.test(RATERUN, RATEMAX)

      Welch Two Sample t-test

data:  RATERUN and RATEMAX
t = -1.6719, df = 59.26, p-value = 0.09982
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -9.0704610  0.8123965
sample estimates:
mean of x mean of y
 169.6452  173.7742


> mod2 = lm(OXYGEN ~AGE +RUNTIME + +RATERUN, fitness)
> summary(mod2)

Call:
lm(formula = OXYGEN ~ AGE + RUNTIME + +RATERUN, data = fitness)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 111.71996   10.24018  10.910 2.12e-11 ***
AGE          -0.25621    0.09628  -2.661   0.0129 *
RUNTIME      -2.82577    0.35846  -7.883 1.78e-08 ***
RATERUN      -0.13095    0.05062  -2.587   0.0154 *
---

Residual standard error: 2.442 on 27 degrees of freedom
Multiple R-squared:  0.811,   Adjusted R-squared: 0.79
```

After removed RATEMAX from the model, the adjused-R^2 decrease by 0.02 by the model is simpler. Therefore, we conclude that RATEMAX is not significant in predicting oxygen consumption.

Conclusion

In the correlation test, we found that oxygen is strongly negative correlated with run time. The variable selection confirms that these correlations are significant with negative coefficients (-2.83). Thus, significant predictors of oxygen consumption include run time along with age and run pulse.