Multiple Linear Regression Initial Proposal

# PROJECT PROPOSAL:

Sports scientist, Thomas Kruz suggested that grip strength provides valuable information about athlete’s potential, readiness and fatigue. The grip strength depends on many factors like palm length, palm width, age, height, weight, food intake, etc.

We propose to determine multiple linear regression model for the collected data of grip strength, palm length, palm width, age, height and weight. In these data grip strength is a response variable (Yi) while palm length, palm width, age, height and weight are predictor variables (Xi). For this project, 30 people were selected randomly from UTA campus and data was collected.

We have used hand dynamometer to measure grip strength in kg, 30cm ruler to measure palm width and palm length in cm. Height and Weight were asked directly. Sometimes weight was measured using weighing machine in kg and height was measured using simple techniques in cm.

This dataset has been provided by Dr. Victoria Chen as the original proposal dataset was incapable for simple linear regression

References: Kurz T. Science of Sports Training. Stadion Publishing Co. 2001.

## EXPLORATORY DATA ANALYSIS

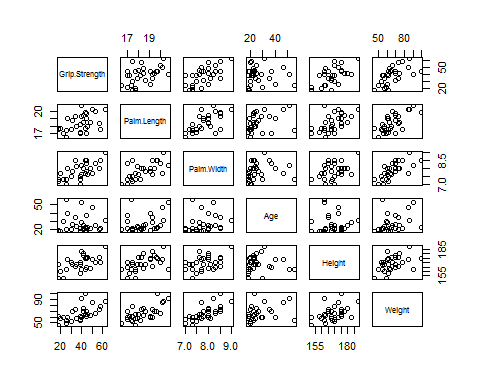
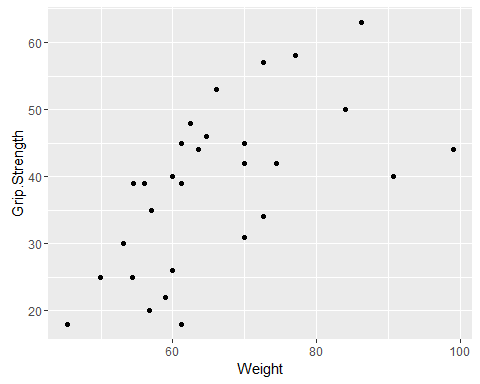
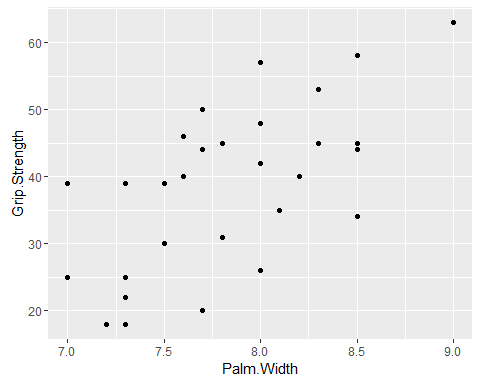
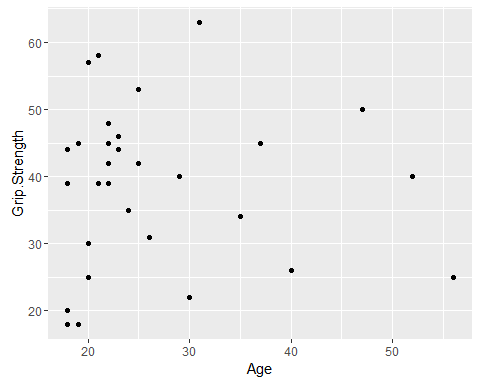


Figure 1

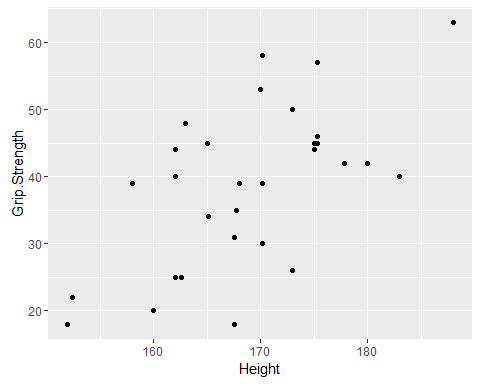
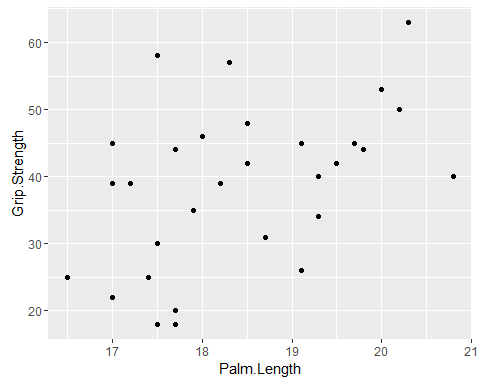
* By looking at the predictor vs Predictor plot we can find a linear increasing trend between
  + PalmLength and Palm Width
  + Palm Length and Height
  + Pam Length and weight
  + Palm Width and Weight
  + Height and Weight
* Response vs Predictor plot can be explained as below



* There is an upward trend in these scatter plots.
* The grip strength tends to increase with increasing palm width.
* The grip strength tends to increase with increasing Weight.
* The outliers are shown in the boxes.
* There appears to be constant residual variation in first figure. However; the residual variation tends to increase in the second figure.



* There is no relationship between “Grip Strength” and “Age” as shown from the above plot.



* There appears to be linear upward trend between Grip Strength and Palm Length. However; there are 8 outliers.
* There appears to be linear upward trend between Grip Strength and Height. There are 2 outliers.

## Grip.Strength Palm.Length Palm.Width Age Height  
## Grip.Strength 1.00000000 0.4630881 0.6395775 -0.01702263 0.60501344  
## Palm.Length 0.46308805 1.0000000 0.6397168 0.35506310 0.53575309  
## Palm.Width 0.63957751 0.6397168 1.0000000 0.15681638 0.46160588  
## Age -0.01702263 0.3550631 0.1568164 1.00000000 -0.01420184  
## Height 0.60501344 0.5357531 0.4616059 -0.01420184 1.00000000  
## Weight 0.60510783 0.6824771 0.7239956 0.32328853 0.44649358  
## Weight  
## Grip.Strength 0.6051078  
## Palm.Length 0.6824771  
## Palm.Width 0.7239956  
## Age 0.3232885

* We can see a very high correlation between Palm Width and Weight (0.72)
* Correlation between Palm Length and Weight (.68) ; Palm Length vs Pam Width (.64) is high but less than 0.7
* All other predictor vs Predictor correlation is than 0.6
* Response variable has high correlation with Pam Width and Height.

### COMPLICATIONS

* We have some outliers which have been shown in the plots. However they have to be tested
* Multicollinearity can be serious between PaLm Width and Weigth(0.72) It is also an issue with Palm Length vs Weight and Palm Length VS Palm Width.
* There is some curvilinearity between Palm Width vs Height, Palm Width vs Age, Palm Width vs Weight.

# PROJECT PROPOSAL: REPORT

Sports scientist, Thomas Kruz suggested that grip strength provides valuable information about athlete’s potential, readiness and fatigue. The grip strength depends on many factors like palm length, palm width, age, height, weight, food intake, etc.

We propose to determine multiple linear regression model for the collected data of grip strength, palm length, palm width, age, height and weight. In these data grip strength is a response variable (Yi) while palm length, palm width, age, height and weight are predictor variables (Xi).For this project, 30 people were selected randomly from UTA campus and data was collected.

We have used hand dynamometer to measure grip strength in kg, 30cm ruler to measure palm width and palm length in cm. Height and Weight were asked directly. Sometimes weight was measured using weighing machine in kg and height was measured using simple techniques in cm.

This dataset has been provided by Dr. Victoria Chen as the original proposal dataset was incapable for simple linear regression

References: Kurz T. Science of Sports Training. Stadion Publishing Co. 2001.

# EXPLORATORY DATA ANALYSIS

## SCATTERPLOT MATRIX

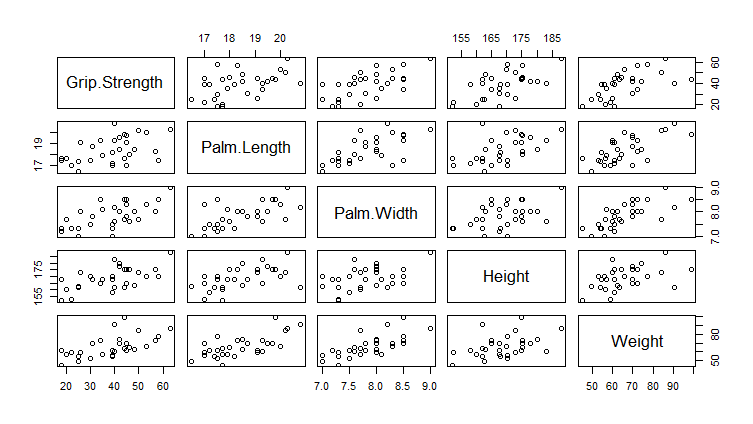


Figure 1 Scatterplot Matrix

The above matrix is generated on R. It gives us the relationship between all the variables. In the first row we can see outcome vs predictor plots; we can see a linear increasing trend between “Grip Strength” Vs“Palm Length” and “Palm Width”. Same can be said for height which is having a x-y outlier. We can see a curvilinear trend between Grip Strength and Weight.

Next we arrive 2nd row to X-X plots between Palm Length and other predictor variables; First and second plots shows an increasing linear trend; there is a Y outlier in the second plot. The third plot shows a slight curvature at the end.

In the third row, 4th column; we can see a linear trend between Palm Width and Height and a curvilinear trend between Palm Width and Weight.

For the fourth row, we have curvilinear relationship between Height and Weight.

There are no X outliers in any of the predictor Vs Predictor plot.

## CORRELATION MATRIX

## Grip.StrengthPalm.LengthPalm.Width Height Weight  
## Grip.Strength 1.0000000 0.4630881 0.6395775 0.6050134 0.6051078  
## Palm.Length 0.4630881 1.0000000 0.6397168 0.5357531 0.6824771  
## Palm.Width 0.6395775 0.6397168 1.0000000 0.4616059 0.7239956  
## Height 0.6050134 0.5357531 0.4616059 1.0000000 0.4464936  
## Weight 0.6051078 0.6824771 0.7239956 0.4464936 1.0000000

Correlation Matrix is generated in R. Since it is symmetric; all diagonal elements are 1. We want to have higher correlations between our outcome Grip Strength and other predictor variables; which are all above 0.6 other than Palm Length (0.46).   
We want to have lower correlations between our X-X plots. We have moderate correlation between Palm Length Vs Height and Height Vs Weight of 0.53 and 0.45 respectively. We have high correlations between Palm Length vs Palm Width and Weight of 0.64 and 0.68 respectively. We have very high correlation and probably multicollinearity problem between Palm Width Vs Height (0.72).

# PRELIMINARY MLR MODEL ANALYSIS

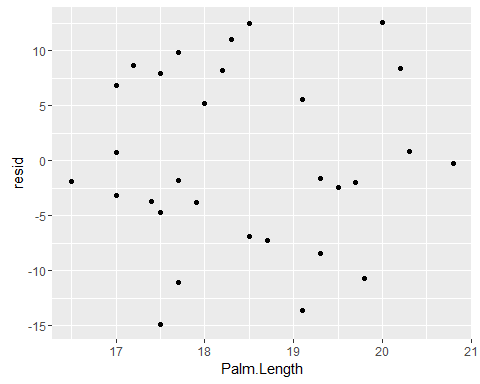
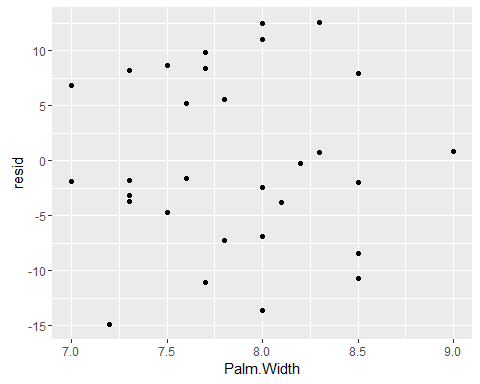
Considering our correlation matrix and scatterplot matrix; we will base our preliminary model with Grip Strength as our outcome and the rest as our predictors.  
 Grip Strength (denoted by Y) as follows

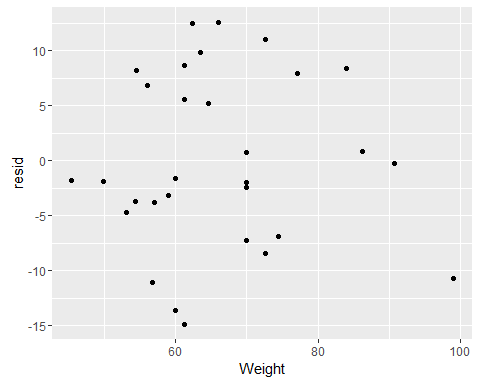
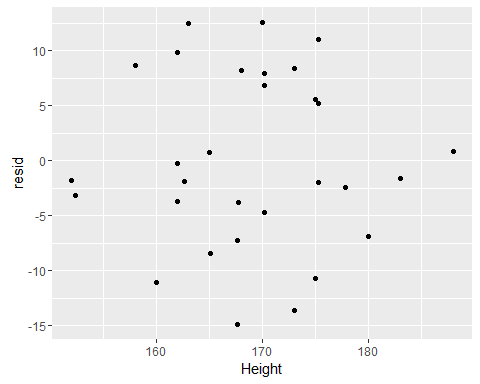
Where x1= Palm Length; x2= Palm Width; x3= Height; x4= Weight; = Regression Coefficients; = Vertical variation between observed and fitted value.

We will check the model assumptions which are as follows

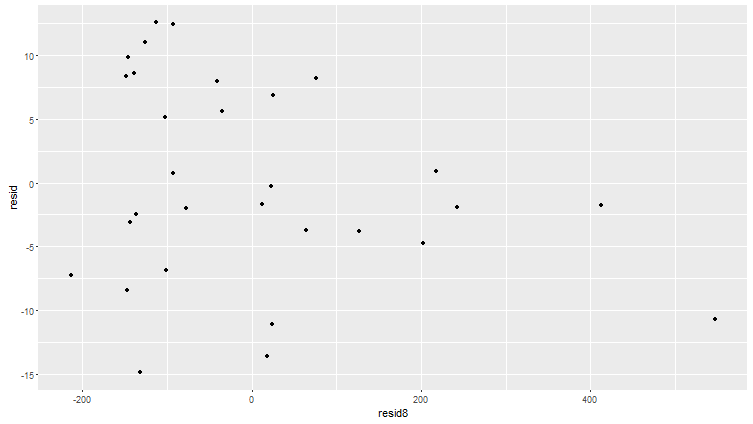
1. Current Model form is reasonable
2. Residuals have constant variance.
3. Model is uncorrelated
4. Residuals have normality.

We will now plot residuals vs each predictor variable using R to check the first assumption that the model form is ok or not.



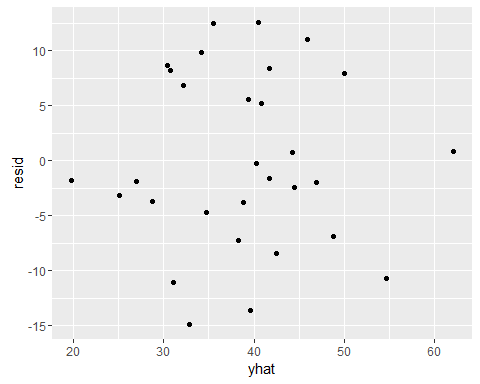


Above are the plots for residuals Vs Predictors. We can see that there is a slight curvature (Inverted U shape) between Residual Vs Weight plot. In all the other plots; there are no curvature between residual vs predictor variables(Palm length, Palm width, Height) and are randomly scattered. We can say that the model form is not reasonable; there must be a square term of weight that has to be added in this model. To be definite we will plot a partial regression plot between residual of our preliminary model e(y|x1,x2,x3,x4) and e(x4^2| x1,x2,x3,x4), where x1= Palm Length; x2= Palm Width, x3= Height, x4= Weight we get,



Here “resid” is e(y|x1,x2,x3,x4) and resid8 is e(x4^2| x1,x2,x3,x4). We can see a downward trend suggesting that weight^2 should be added to the preliminary model.

To check for constant variance:



Residuals vs fitted values

Above plot is between residuals and fitted values. as we can see the inverted U curvature which has been confirmed between residual Vs Weight plot. However; to test if this curvature is significant or not we will perform modified Levine test to find if variance is constant or not.

## Modified-levene test:

We will first divide the data into two sets. then calculate absolute deviations of the residual around the group medians. Conduct two-sample t-test.

Two-sample T for absolutede

Group N MeanStDev SE Mean

1 15 6.56 4.95 1.3

2 15 6.16 4.65 1.2

Difference = μ (1) - μ (2)

Estimate for difference: 0.40

95% CI for difference: (-3.20, 4.00)

T-Test of difference = 0 (vs ≠): T-Value = 0.23 P-Value = 0.820 DF = 27



Box plot of Absolute deviation

Above is the box plot for the Absolute deviations of residuals around their group medians.  
Null Hypothesis H0: Variance is equal   
Alternative Hypothesis H1: Variance is unequal  
P-value : 0.820  
𝛼 : 0.05  
P-value >𝛼  
We fail to reject H0 , So we can conclude that Variance is equal and hence the error variance is constant.

### NORMALITY



Normal probability plot

Above is the normal probability plot for residuals. We can see from the plot that most of the points are in line with the normality line. In right hand side the tail is very slightly longer than the normal line and in left hand side the tail is shorter.

Result of Normality test in Minitab

Null Hypothesis H0: Normality is ok  
Alternative Hypothesis H1: Normality is not ok  
P-value :0.265  
𝛼 : 0.05  
P-value >𝛼  
We fail to reject H0 , So we can conclude that Normality is ok. We are 95% Confident that the Normality is ok.

To check if Model is Uncorrelated:

As we have collected data randomly in different time period, Time sequence plot cannot be used in verifying model adequacy. We will assume that the data is uncorrelated.

### Benferroni outlier test:



FDiagnostics of Observation

We will use Bonferroni test to check for Y-outliers. As shown in the above diagram we have calculated Studentised residual, the deleted residual, The Studentized deleted residual. We will use significance level of 0.05. No.of Parameters is 5 and No. of observation is 30.

Bonferroni Cutoff: t (1-𝛼/2𝑛, n-p-1)

𝛼 = 0.05 p = 5 n = 30

t-table cutoff value :3.26

When we compare 3.26 with the Studentized deleted residuals we can see that no value is bigger than than 3.26. hence we can conclude that there is no Y outlier.

X-Outlier test:

We will use the leverage value to perform x-Outlying test. If the Leverage value is more than 2p/n value which is 0.333 for our case then we can say that it is X outlying observation. Here we see that h27,27  = 0.4339. X is an outlying observation.

We will now calculate the if it has any influence on Regression.

Influence on Fitted values:  
DFITS27 = -0.368  
DFITS27 < 1  
We can conclude that Observation 27 is not influential.

Cook's Distance:  
D27 = e272/p\*MSE  
D27 = -0.26952/5\*72.22  
D27 = 0.00020105  
F( 0.50,5,25) = 0.894  
D27 < F( 0.50, 5 ,25)  
We can Conclude that observation 27 is not Influential.

By analysing all the above mentioned data we have concluded that:

1. As there are no curvature in the plots of residual vs predictor variables ( Palm length, Palm width, Height, Weight) and are randomly scattered, We can say that the model form is ok
2. After plotting residuals and fitted values, It had no funnel shape and curvature. Hence we can say that variance is constant.
3. After performing Normality test we concluded that Normality is ok.
4. The data were uncorrelated.

Analysis of Variance:

Analysis of Variance

Source DF AdjSS Adj MS F-Value P-Value

Regression 4 2357.79 589.45 8.16 0.000

x1 1 60.31 60.31 0.83 0.370

x2 1 231.08 231.08 3.20 0.086

x3 1 477.02 477.02 6.60 0.017

x4 1 133.66 133.66 1.85 0.186

Error 25 1805.58 72.22

Total 29 4163.37

Model Summary

S R-sq R-sq(adj) R-sq(pred)

8.49843 56.63% 49.69% 39.50%

Coefficients

Term Coef SECoef T-Value P-Value VIF

Constant -112.1 40.4 -2.78 0.010

x1 -1.87 2.04 -0.91 0.370 2.26

x2 8.64 4.83 1.79 0.086 2.34

x3 0.587 0.228 2.57 0.017 1.45

x4 0.279 0.205 1.36 0.186 2.54

Regression Equation

y = -112.1 - 1.87 x1 + 8.64 x2 + 0.587 x3 + 0.279 x4

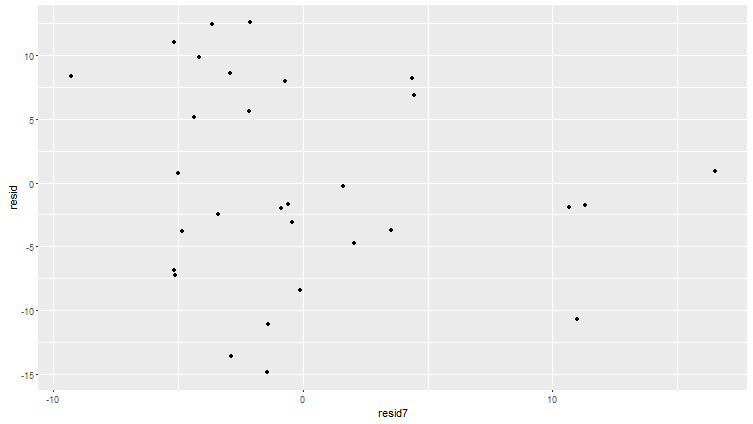
Above is the output Of ANOVA table from minitab.  
Null Hypothesis H0:β1....βk= 0  
Alternative Hypothesis H1: at least one βk not 0  
P-Value = 0.0000, 𝛼 = 0.05 Here 0.0000<0.05, We reject H0 and conclude that Multiple Linear Regression model is Significant.

From the ANOVA table we can see that MSE=72.22, SSE=1805.58, MSR=589.45, SSR=235779. The R2 value is 56.63% Which represents the variability explained in the model.

**EXPLORATION OF INTERACTION TERM:**

Interaction terms can be obtained by Multiplying Predictor variables with each other. We can find out the effect by plotting the interaction terms and there residuals. We have to first generate residuals of predictor-predictor variable and then plot with the residuals.

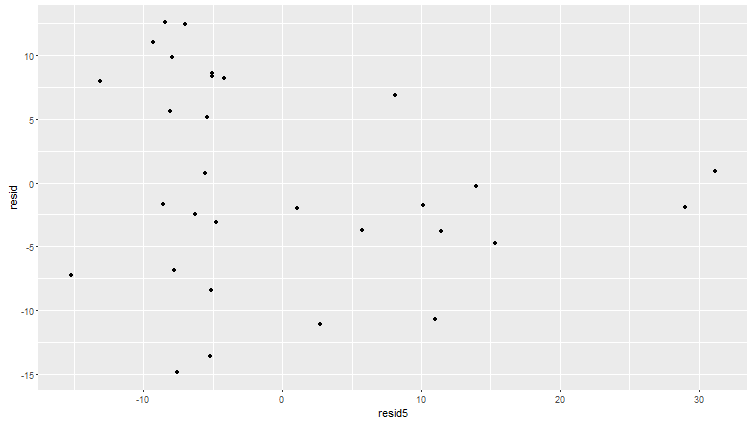
1) Interaction term obtained by multiplying Palm width and Weight.



resid = e(Y| X1, X2, X3, X4), resid7 = e(X2X4| X1, X2, X3, X4)

After analysing the plot between the residuals, we can see that there is a downward trend in the plot. We conclude that this model will have effect on our regression plot and we need to add x2x4.

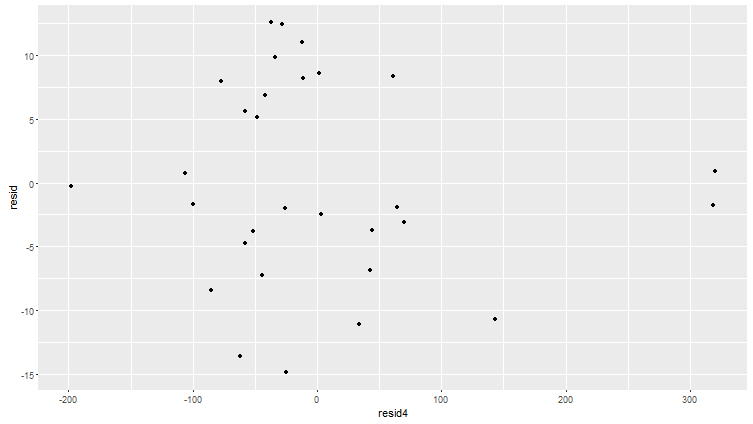
2) Interaction term obtained by multiplying Weight and palm length.



resid = e(Y| X1, X2, X3, X4), resid5 = e(X1X4| X1, X2, X3, X4)

After analysing the plot between the residuals, we can see that there is a downward trend in the plot. We conclude that this model will have effect on our regression plot.

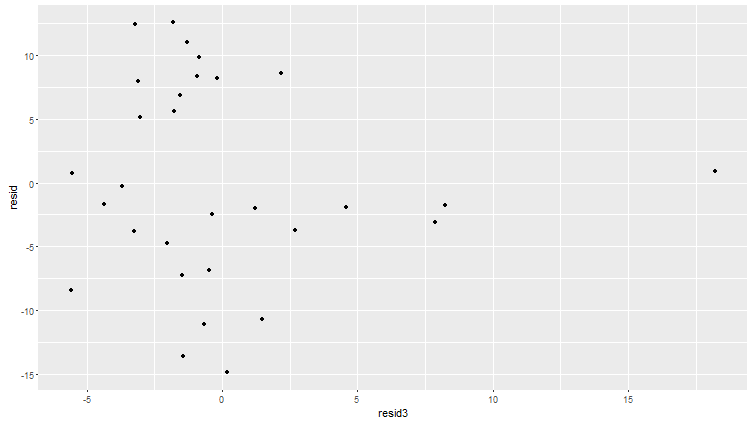
3) Interaction term obtained by multiplying Weight and Height.



resid = e(Y| X1, X2, X3, X4), resid 4= e(X3X4| X1, X2, X3, X4)

After analysing the plot between the residuals, we can see that there is a Downward trend in the plot. We conclude that this model will have effect on our regression plot.

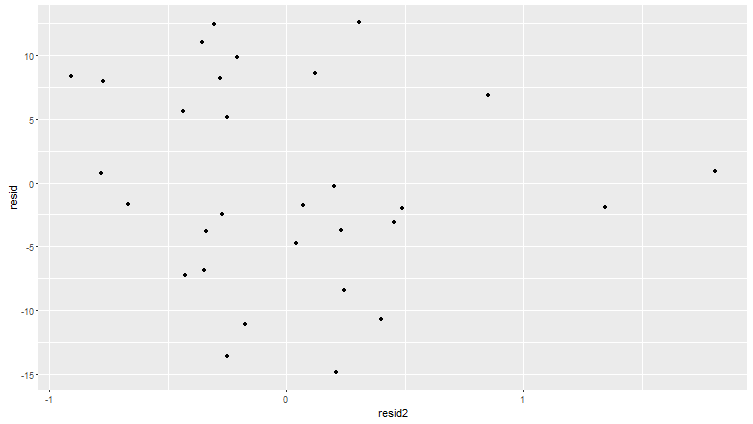
4) Interaction term obtained by multiplying Palm weight and Height.



resid = e(Y| X1, X2, X3, X4), resid3 = e(X2X3| X1, X2, X3, X4)

After analysing the plot between the residuals, we can see that there is a no trend in the plot. We conclude that this model will have no effect on our regression plot.

5) Interaction term obtained by multiplying Palm width and Palm length.



resid = e(Y| X1, X2, X3, X4), resid2 = e(X1X2|X1, X2, X3, X4)

After analysing the plot between the residuals, we can see that there is a Downward trend in the plot. We conclude that this model will have effect on our regression plot.

5) Interaction term obtained by multiplying height and Palm length.



resid = e(Y| X1, X2, X3, X4), resid6 = e(X1X3| X1, X2, X3, X4)

After analysing the plot between the residuals, we can see that there is a Downward trend in the plot. We conclude that this model will have effect on our regression plot.

As we can see from the interaction model plot , the interaction terms between Palm height, Palm width, Height, Weight have effect on the regression model as they show downward trend. So we will add this interaction term to our regression model and find the correlation.

Below is the result for correlation between the variables from R studio. As we can see from the result that the corelation is significantly high for some of the variables. We will now check if the co-relation is reduced when we use the standarized interaction model by centering mean to 0 and variance 1.

Grip.StrengthPalm.LengthPalm.Width Height Weight x1x2 x1x3

Grip.Strength 1.0000000 0.4630881 0.6395775 0.6050134 0.6051078 0.6047832 0.5980477

Palm.Length 0.4630881 1.0000000 0.6397168 0.5357531 0.6824771 0.9061433 0.9039837

Palm.Width 0.6395775 0.6397168 1.0000000 0.4616059 0.7239956 0.9035936 0.6400963

Height 0.6050134 0.5357531 0.4616059 1.0000000 0.4464936 0.5503510 0.8444718

Weight 0.6051078 0.6824771 0.7239956 0.4464936 1.0000000 0.7774287 0.6568859

x1x2 0.6047832 0.9061433 0.9035936 0.5503510 0.7774287 1.0000000 0.8533643

x1x3 0.5980477 0.9039837 0.6400963 0.8444718 0.6568859 0.8533643 1.0000000

x2x4 0.6420363 0.7029181 0.8351527 0.4735227 0.9826124 0.8507405 0.6847701

x4x1 0.5896131 0.8110231 0.7331190 0.4842377 0.9796581 0.8546356 0.7578728

x2x4 x4x1

Grip.Strength 0.6420363 0.5896131

Palm.Length 0.7029181 0.8110231

Palm.Width 0.8351527 0.7331190

Height 0.4735227 0.4842377

Weight 0.9826124 0.9796581

x1x2 0.8507405 0.8546356

x1x3 0.6847701 0.7578728

x2x4 1.0000000 0.9687257

x4x1 0.9687257 1.0000000

Standardized correlation:

Below is the result for standardized correlation. We can see from the below result that there is no change in the correlation when we change the Interaction term.

Grip.Strength Palm.Length Palm.Width Height Weight stdx1x2 stdx1x3

Grip.Strength 1.0000000 0.4630881 0.6395775 0.6050134 0.6051078 0.6047832 0.5980477

Palm.Length 0.4630881 1.0000000 0.6397168 0.5357531 0.6824771 0.9061433 0.9039837

Palm.Width 0.6395775 0.6397168 1.0000000 0.4616059 0.7239956 0.9035936 0.6400963

Height 0.6050134 0.5357531 0.4616059 1.0000000 0.4464936 0.5503510 0.8444718

Weight 0.6051078 0.6824771 0.7239956 0.4464936 1.0000000 0.7774287 0.6568859

stdx1x2 0.6047832 0.9061433 0.9035936 0.5503510 0.7774287 1.0000000 0.8533643

stdx1x3 0.5980477 0.9039837 0.6400963 0.8444718 0.6568859 0.8533643 1.0000000

stdx2x4 0.6420363 0.7029181 0.8351527 0.4735227 0.9826124 0.8507405 0.6847701

stdx4x1 0.5896131 0.8110231 0.7331190 0.4842377 0.9796581 0.8546356 0.7578728

stdx2x4 stdx4x1

Grip.Strength 0.6420363 0.5896131

Palm.Length 0.7029181 0.8110231

Palm.Width 0.8351527 0.7331190

Height 0.4735227 0.4842377

Weight 0.9826124 0.9796581

stdx1x2 0.8507405 0.8546356

stdx1x3 0.6847701 0.7578728

stdx2x4 1.0000000 0.9687257

stdx4x1 0.9687257 1.0000000

# MODEL SEARCH

We will try to find the best 2 models using backward deletion, best subset and stepwise regression methods. The best models have the predictor variables having p values less than 0.05 and will not have serious multicollinearity. For our basis we will add the variables(interaction terms) to our preliminary model. Please note we have not added stdx2x3, stdx3x4 as they don’t show any trend in our previous analysis plots.

## STEPWISE REGRESSION

This method rejects the variables by checking their p values. In order to be significant in our model the p value of each predictor shall be less than 0.05.

After performing stepwise regression our best model is where

Y= Grip Strength, X2= Palm Width; X3=Height  
R output is given below

Call:  
## lm(formula = Grip.Strength ~ Palm.Width + Height, data = data)  
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -142.9995 33.9854 -4.208 0.000255 \*\*\*  
## Palm.Width 10.9857 3.5647 3.082 0.004698 \*\*   
## Height 0.5663 0.2137 2.650 0.013304 \*   
## ---  
## Residual standard error: 8.504 on 27 degrees of freedom  
## Multiple R-squared: 0.531, Adjusted R-squared: 0.4963   
## F-statistic: 15.29 on 2 and 27 DF, p-value: 3.637e-05  
 VIF= 2.1321, AIC= 131.27

## BACKWARD DELETION MODEL

In this method we will take all the predictor variables with interaction terms which shows trend with Grip Strength. At each iteration predictors which are not significant at α= 0.1 is removed. This process continues till all the predictors satisfy this condition.  
For our model we start with

Start: AIC=139.67

Grip.Strength ~ Palm.Length + Palm.Width + Height + Weight^2 +

stdx1x2 + stdx4x1 + stdx2x4 + stdx1x3

Df Sum of Sq RSS AIC

- Palm.Length 1 0.001 1731.8 137.67

- stdx1x3 1 0.188 1732.0 137.68

- Palm.Width 1 0.209 1732.0 137.68

- Height 1 0.545 1732.4 137.68

- stdx1x2 1 0.955 1732.8 137.69

- stdx2x4 1 3.803 1735.6 137.74

- stdx4x1 1 16.642 1748.5 137.96

- Weight 1 43.889 1775.7 138.42

<none> 1731.8 139.67

We end with

Step: AIC=132.13

Grip.Strength ~ Height + Weight + stdx1x2 + stdx4x1

Df Sum of Sq RSS AIC

<none> 1758.5 132.13

- stdx1x2 1 239.96 1998.4 133.97

- stdx4x1 1 267.98 2026.4 134.38

- Weight 1 369.64 2128.1 135.85

- Height 1 416.79 2175.2 136.51

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -170.3229 63.0114 -2.703 0.0122 \*

Height 0.5455 0.2241 2.434 0.0224 \*

Weight 1.7698 0.7720 2.292 0.0306 \*

stdx1x2 7.0051 3.7927 1.847 0.0766 .

stdx4x1 -22.4126 11.4825 -1.952 0.0622 .

Multiple R-squared: 0.5776, Adjusted R-squared: 0.5101

VIF= 2.3674

## BEST SUBSET REGRESSION

We use the same model as we used in the backward deletion method. This method selects the best 2 models using 4 criteria.

* High R2
* High adjusted R2
* Low Mallow (Cp) values
* Low values of AIC

Below output :

First column represents the number of variables in the model. TRUE represents variable is included in the model. FALSE represents variable is not included in the model. Correspondingly all the values of cp, r2 and adjusted r2 is represented. Ex:- first value of cp/r2/adjr2 is for the first row

$which

Palm.Length Palm.Width Height Weight stdx1x2 stdx1x3 stdx2x4 stdx4x1

1 FALSE FALSE FALSE FALSE FALSE FALSE TRUE FALSE

1 FALSE TRUE FALSE FALSE FALSE FALSE FALSE FALSE

2 FALSE TRUE TRUE FALSE FALSE FALSE FALSE FALSE

2 FALSE FALSE TRUE FALSE FALSE FALSE TRUE FALSE

3 FALSE FALSE TRUE FALSE FALSE FALSE TRUE TRUE

3 FALSE TRUE TRUE TRUE FALSE FALSE FALSE FALSE

4 FALSE FALSE FALSE TRUE TRUE TRUE FALSE TRUE

4 FALSE TRUE FALSE TRUE FALSE TRUE FALSE TRUE

5 FALSE FALSE TRUE TRUE TRUE FALSE TRUE TRUE

5 FALSE TRUE FALSE TRUE FALSE TRUE TRUE TRUE

6 FALSE FALSE TRUE TRUE TRUE TRUE TRUE TRUE

6 TRUE FALSE TRUE TRUE TRUE FALSE TRUE TRUE

7 FALSE TRUE TRUE TRUE TRUE TRUE TRUE TRUE

7 TRUE TRUE TRUE TRUE TRUE FALSE TRUE TRUE

8 TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE

$adjr2

[1] 0.3912182 0.3879544 0.4962682 0.4941047 0.5074274 0.5001203 0.5151507 0.5156412 0.4972644

[10] 0.4969853 0.4754504 0.4754252 0.4516751 0.4516160 0.4255648

$r2

[1] 0.4122107 0.4090594 0.5310084 0.5289941 0.5583832 0.5518320 0.5820265 0.5807252 0.5839429

[10] 0.5837119 0.5839779 0.5839579 0.5840294 0.5839845 0.5840297

$Cp

[1] 3.6741728 3.8332632 -0.3232534 -0.2215638 0.2947474 0.6254800 1.1011276 1.1668230

[9] 3.0043780 3.0160392 5.0026142 5.0036227 7.0000121 7.0022775 9.0000000

Considering adjusted R^2 we can see it decreases in 5 variable model. The highlighted ones are the highest in 4 predictor model suggesting we should not include any more predictors in our model. Our R^2 values increases from left to right suggesting 8 predictor model should be taken. However there is only a marginally decline in R^2 in 4 predictor models (0.2). Our Cp values decrease till 2 predictor model suggesting 2 predictor model can be taken.

# MODEL SELECT

Based on the intersection of all 3 methods, our 2 final models are:-

### Model1 (Best)

Grip.Strength ~ Palm.Width + Height

### Model2 (Second Best)

Grip.Strength ~ Palm.Width + Weight^2 + stdx1x3 + stdx4x1

### For Model 1

Let us see the coefficients for model 1

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -142.9995 33.9854 -4.208 0.000255 \*\*\*

Palm.Width 10.9857 3.5647 3.082 0.004698 \*\*

Height 0.5663 0.2137 2.650 0.013304 \*

---

Residual standard error: 8.504 on 27 degrees of freedom

Multiple R-squared: 0.531, Adjusted R-squared: 0.4963

F-statistic: 15.29 on 2 and 27 DF, p-value: 3.637e-05  
VIF= 2.1321, AIC= 131.27

We can see all the predictors have t value less than 0.05 telling us the predictors are significant

Hence the model is

Grip strength = -143.0 + 10.99 Palm Width + 0.566 Height

We selected this model because

* This is the best model after 4 predictor models. The 4 predictor models are our second choice due to one of the predictors have p value greater than 0.05
* Cp is the lowest for this model. Adjusted R^2 and R2 are 0.50 and 0.55 respectively.
* There is correlation of 0.46 which is one of the lowest in the dataset, hence no multicollinearity problem. Besides all the predictors have p value less than 0.05.

We will check the model assumptions which are as follows

1. Current Model form is reasonable
2. Residuals have constant variance.
3. Model is uncorrelated
4. Residuals have normality.

We will now plot residuals v s each predictor variable using R to check the first assumption that the model form is ok or not.

To check for model adequacy:



From the above two plots between residuals and predictors we can see that there is no pattern, points are randomly scattered. We can conclude that the model form is adequate.

To Check for constant Variance:



Residuals vs fitted values

As we can see from the above plot that there is no pattern in the plot, Hence we can conclude that the variance is constant.

To check for Normality:



Above is the normal probability plot for residuals. We can see from the plot that most of the points are in line with the normality line. In right hand side the tail is very slightly longer than the normal line and in left hand side the tail is shorter.

Result of Normality test in Minitab

Null Hypothesis H0: Normality is ok

Alternative Hypothesis H1: Normality is not ok  
P-value :0.106  
𝛼 : 0.05  
P-value >𝛼  
We fail to reject H0 , So we can conclude that Normality is ok. We are 95% Confident that the Normality is ok.

To check if Model is Uncorrelated:

As we have collected data randomly in different time period, Time sequence plot cannot be used in verifying model adequacy. We will assume that the data is uncorrelated.

Diagnostics:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| FITS | RESI | SRES | TRES | HI | COOK | DFIT |
| 37.19629 | 10.80371 | 1.316518 | 1.335483 | 0.068793 | 0.04268 | 0.362983 |
| 44.16205 | 12.83795 | 1.552044 | 1.595883 | 0.053898 | 0.045743 | 0.380908 |
| 30.28814 | 8.711863 | 1.123956 | 1.12969 | 0.169237 | 0.085782 | 0.50988 |
| 49.65488 | -4.65488 | -0.57529 | -0.56803 | 0.094703 | 0.011541 | -0.18372 |
| 46.76664 | 11.23336 | 1.394332 | 1.42036 | 0.102486 | 0.074001 | 0.479966 |
| 46.82376 | -4.82376 | -0.59748 | -0.59022 | 0.098679 | 0.013028 | -0.19529 |
| 62.34 | 0.660004 | 0.091786 | 0.090084 | 0.285022 | 0.001119 | 0.056878 |
| 44.45624 | 8.543758 | 1.039726 | 1.04135 | 0.066286 | 0.025581 | 0.27746 |
| 49.48498 | -5.48498 | -0.67772 | -0.67078 | 0.094252 | 0.015932 | -0.21638 |
| 39.76778 | 6.232222 | 0.765938 | 0.759921 | 0.084512 | 0.018052 | 0.230888 |
| 45.57785 | -3.57785 | -0.43727 | -0.43062 | 0.074244 | 0.005111 | -0.12195 |
| 40.95657 | -5.95657 | -0.71829 | -0.71169 | 0.049062 | 0.008873 | -0.16165 |
| 25.98409 | -0.98409 | -0.12413 | -0.12185 | 0.130923 | 0.000774 | -0.04729 |
| 35.78097 | -5.78097 | -0.70083 | -0.69407 | 0.059129 | 0.010289 | -0.174 |
| 31.01284 | -13.0128 | -1.61153 | -1.66343 | 0.098389 | 0.094468 | -0.5495 |
| 41.62463 | 3.375365 | 0.417772 | 0.411294 | 0.097353 | 0.006275 | 0.135073 |
| 43.8784 | -9.8784 | -1.25489 | -1.269 | 0.143129 | 0.087681 | -0.51864 |
| 23.50332 | -1.50332 | -0.19398 | -0.19049 | 0.169522 | 0.00256 | -0.08606 |
| 28.87185 | 10.12815 | 1.250148 | 1.263904 | 0.092407 | 0.053041 | 0.403292 |
| 32.20163 | -12.2016 | -1.49141 | -1.52782 | 0.074464 | 0.059652 | -0.43336 |
| 23.27679 | -5.27679 | -0.68351 | -0.67661 | 0.175849 | 0.033228 | -0.31254 |
| 32.33793 | 6.662069 | 0.816957 | 0.811781 | 0.080453 | 0.019465 | 0.240117 |
| 41.79502 | 3.204985 | 0.388656 | 0.382462 | 0.059682 | 0.003196 | 0.096355 |
| 33.33427 | 10.66573 | 1.291747 | 1.308683 | 0.057285 | 0.033798 | 0.3226 |
| 42.85951 | -16.8595 | -2.0257 | -2.15862 | 0.042155 | 0.060197 | -0.45285 |
| 28.94 | -3.94 | -0.4826 | -0.47563 | 0.078325 | 0.006597 | -0.13865 |
| 38.8271 | 1.172897 | 0.146203 | 0.143527 | 0.110057 | 0.000881 | 0.050473 |
| 39.56381 | 10.43619 | 1.261172 | 1.275744 | 0.053132 | 0.029751 | 0.302202 |
| 44.12845 | -4.12845 | -0.5436 | -0.53638 | 0.202426 | 0.024999 | -0.27022 |
| 37.60424 | -6.60424 | -0.79021 | -0.78457 | 0.034147 | 0.007359 | -0.14752 |

There are no x outliers as Hi<0.4

Final Multiple regression model:

After the above Analysis of different regression model, We came to the conclusion that the best model among all different models we have considered is:

Grip strength (y) = -143.0 + 10.99 Palm Width + 0.566 Height

We will use F-test to find out whether our best final model is significant or not.

Below is the output of regression on minitab:

**Regression Analysis: y versus Palm Width, Height**

Analysis of Variance

Source DF Adj SS Adj MS F-Value P-Value

Regression 2 2210.8 1105.39 15.29 0.000

Palm Width 1 686.8 686.82 9.50 0.005

Height 1 507.7 507.72 7.02 0.013

Error 27 1952.6 72.32

Total 29 4163.4

Confidence Interval:  
B= t (1-𝛼/2𝑛, n-p-1) = 2.246  
Two-sided Interval : bk ± B\*s{bk}  
b1= 10.99 b2=0.566 s{b1}=3.56 s{b1}=0.214  
C.I for Palm Width: ( 2.994, 18.98576)  
C.I for Height: (0.0853, 1.046644)  
We will now obtain 95% CI , PI ,CB for Palm width of 9.5 and height of 170.

### For Model 2

Let us see the coefficients for model 2

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -158.3856 60.3271 -2.625 0.0146 \*

Palm.Width 6.8110 4.7759 1.426 0.1662

Weight 2.1752 0.8538 2.548 0.0174 \*

stdx1x3 8.9511 3.2821 2.727 0.0115 \*

stdx4x1 -28.3795 11.8763 -2.390 0.0247 \*

Residual standard error: 8.356 on 25 degrees of freedom

Multiple R-squared: 0.5807, Adjusted R-squared: 0.5136

F-statistic: 8.657 on 4 and 25 DF, p-value: 0.0001578

We can see Palm.width has a t value greater than 0.05 suggesting it is not a good predictor. However ,this model is the second best model due to the following reason

* Adjusted R^2 is the highest among all possible models. Any increase in the number of predictors decreases its value, thus all 6,7,8 predictor models are eliminated.
* We rejected the second best 4 model predictor model Grip.Strength~Weight^2+stdx1x3+stdx4x1+stdx1x2

On the basis that there exist a high collinearity between stdx1x3 vs stdx1x2 (0.85) and stdx1x3 vs Weight(0.9), there can be a serious multicollinearity problem in the above model

* We rejected 1 predictor models on the basis of Cp and R^2

Model Selection:

Model2: Grip strength = -158.38 + 6.81 Palm width + 2.17Weightsq + 8.95 STD X1X3 – 28.37 STD X4X1

We will check the model assumptions which are as follows

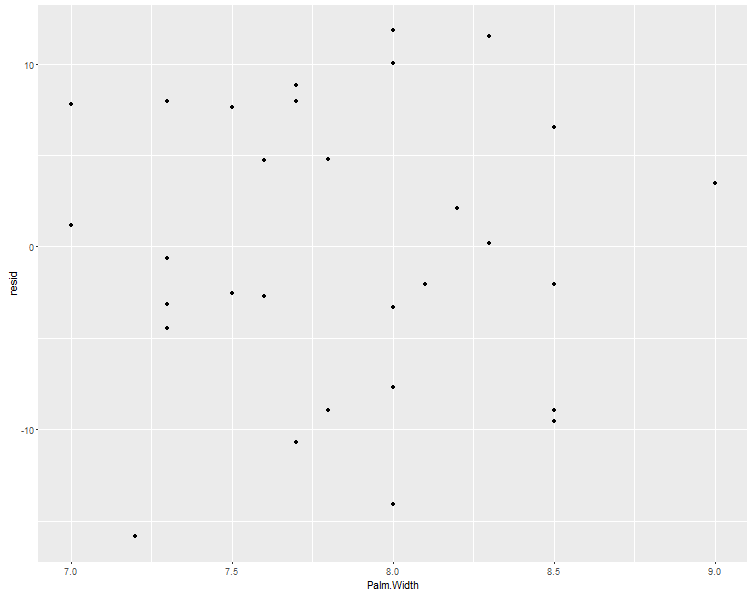
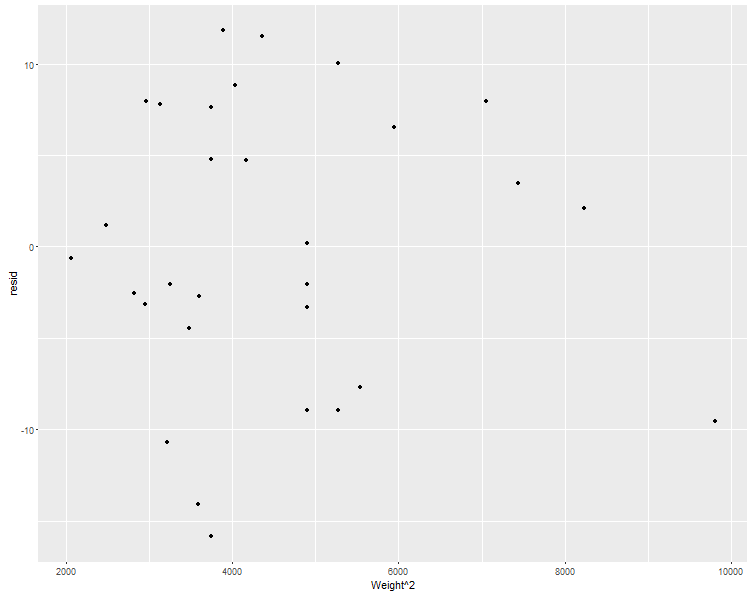
1. Current Model form is reasonable
2. Residuals have constant variance.
3. Model is uncorrelated
4. Residuals have normality.

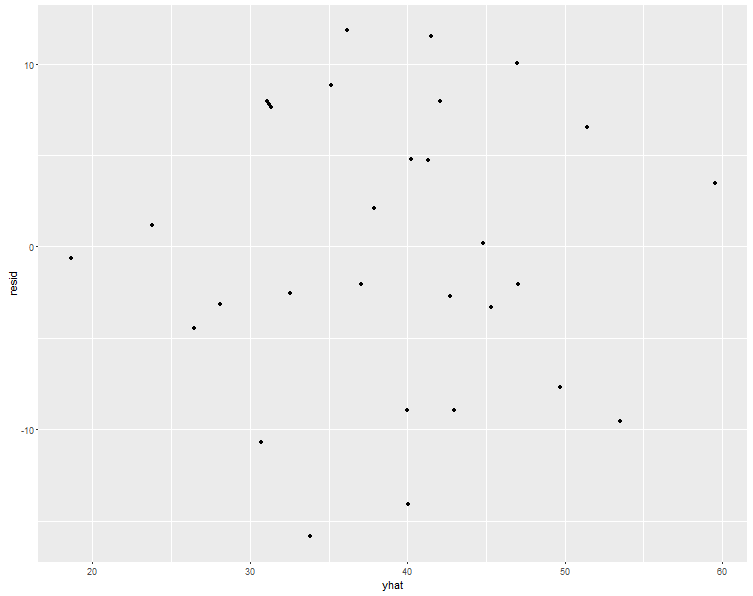
We will now plot residuals v s each predictor variable using R to check the first assumption that the model form is ok or not.

To check for model adequacy:

From the below two plots between residuals and predictors we can see that there is no pattern, points are randomly scattered. We can conclude that the model form is adequate.

To Check for constant Varaince:





Residuals vs fitted values

Above plot is between residuals and fitted values. as we can see the inverted U curvature which has been confirmed between residual Vs Weight plot. However; to test if this curvature is significant or not we will perform modified Levine test to find if variance is constant or not.

To check for Normality:



Above is the normal probability plot for residuals. We can see from the plot that most of the points are in line with the normality line. In right hand side the tail is very slightly longer than the normal line and in left hand side the tail is shorter.

Result of Normality test in Minitab

Null Hypothesis H0: Normality is ok

Alternative Hypothesis H1: Normality is not ok

P-value :0.112

𝛼 : 0.05

P-value >𝛼

We fail to reject H0 , So we can conclude that Normality is ok. We are 95% Confident that the Normality is ok.

To check if Model is Uncorrelated:

As we have collected data randomly in different time period, Time sequence plot cannot be used in verifying model adequacy. We will assume that the data is uncorrelated.

Dignostics:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| FITS\_1 | RESI\_1 | SRES | TRES | HI | COOK | DFIT |
| 39.4632 | 8.536804 | 0.97278 | 0.971694 | 0.075078 | 0.015363 | 0.276842 |
| 45.04996 | 11.95004 | 1.375115 | 1.401373 | 0.093006 | 0.038781 | 0.448752 |
| 26.10717 | 12.89283 | 1.538658 | 1.584459 | 0.15675 | 0.088017 | 0.683136 |
| 45.8943 | -0.8943 | -0.10518 | -0.10308 | 0.131723 | 0.000336 | -0.04015 |
| 53.35669 | 4.643311 | 0.568733 | 0.560882 | 0.19946 | 0.016118 | 0.279968 |
| 45.84896 | -3.84896 | -0.44503 | -0.43778 | 0.10164 | 0.004482 | -0.14725 |
| 52.56822 | 10.43178 | 1.910629 | 2.025761 | 0.64198 | 1.309171 | 2.712658 |
| 43.32619 | 9.673809 | 1.122263 | 1.128379 | 0.10762 | 0.030378 | 0.391856 |
| 54.54779 | -10.5478 | -1.48645 | -1.52539 | 0.395266 | 0.28884 | -1.23323 |
| 37.62642 | 8.37358 | 0.957802 | 0.956158 | 0.082059 | 0.016402 | 0.285882 |
| 43.08561 | -1.08561 | -0.1226 | -0.12016 | 0.058262 | 0.000186 | -0.02989 |
| 36.25092 | -1.25092 | -0.15159 | -0.1486 | 0.182206 | 0.001024 | -0.07014 |
| 20.17723 | 4.822769 | 0.623528 | 0.615737 | 0.281503 | 0.030465 | 0.385411 |
| 28.67426 | 1.325744 | 0.155764 | 0.152691 | 0.12998 | 0.000725 | 0.059018 |
| 32.69727 | -14.6973 | -1.7269 | -1.80296 | 0.130073 | 0.089181 | -0.69717 |
| 47.27 | -2.27 | -0.26228 | -0.25734 | 0.100372 | 0.001535 | -0.08596 |
| 46.22635 | -12.2263 | -1.4563 | -1.49155 | 0.153488 | 0.076909 | -0.63512 |
| 32.62896 | -10.629 | -1.33595 | -1.35835 | 0.239774 | 0.112583 | -0.76285 |
| 35.19675 | 3.80325 | 0.436983 | 0.429798 | 0.090243 | 0.003788 | 0.135366 |
| 33.54558 | -13.5456 | -1.53554 | -1.58092 | 0.065424 | 0.033013 | -0.41828 |
| 24.72267 | -6.72267 | -0.81421 | -0.80855 | 0.181245 | 0.029351 | -0.38042 |
| 30.67322 | 8.326781 | 0.953313 | 0.951507 | 0.083722 | 0.016608 | 0.287619 |
| 38.80857 | 6.191429 | 0.697979 | 0.690639 | 0.054979 | 0.005669 | 0.166582 |
| 38.00473 | 5.995272 | 0.672749 | 0.665206 | 0.046205 | 0.004385 | 0.14641 |
| 39.56132 | -13.5613 | -1.54892 | -1.59614 | 0.079355 | 0.041359 | -0.46861 |
| 28.67662 | -3.67662 | -0.42124 | -0.4142 | 0.085073 | 0.0033 | -0.1263 |
| 40.79936 | -0.79936 | -0.14377 | -0.14093 | 0.628755 | 0.007002 | -0.1834 |
| 42.15642 | 7.843584 | 0.97797 | 0.977083 | 0.227458 | 0.05632 | 0.530177 |
| 38.20249 | 1.79751 | 0.209531 | 0.205478 | 0.116126 | 0.001154 | 0.074479 |
| 41.85277 | -10.8528 | -1.24079 | -1.25497 | 0.081176 | 0.027203 | -0.37302 |

Final Multiple regression model:

After the above Analysis of different regression model, We came to the conclusion that the best model among all different models we have considered is:

Grip strength (y) = -143.0 + 10.99 Palm Width + 0.566 Height

We will use F-test to find out whether our best final model is significant or not.

Below is the output of regression on minitab:

**Regression Analysis: y versus Palm Width, Height**

Analysis of Variance

Source DF Adj SS Adj MS F-Value P-Value

Regression 2 2210.8 1105.39 15.29 0.000

Palm Width 1 686.8 686.82 9.50 0.005

Height 1 507.7 507.72 7.02 0.013

Error 27 1952.6 72.32

Total 29 4163.4

Model Summary

S R-sq R-sq(adj) R-sq(pred)

8.50400 53.10% 49.63% 44.37%

Coefficients

Term Coef SE Coef T-Value P-Value VIF

Constant -143.0 34.0 -4.21 0.000

Palm Width 10.99 3.56 3.08 0.005 1.27

Height 0.566 0.214 2.65 0.013 1.27

Regression Equation

y = -143.0 + 10.99 Palm Width + 0.566 Height

We will now perform F-test to find out whether the regression model is significant or not:

Null hypothesis H0 : β1 = β2 = 0

Alternative hypothesis H1 : Either β1 or β2 not equal to 0

𝛼 = 0.05

P-value < 0.05

We can conclude that the regression is significant.

Confidence Interval:

B= t (1-𝛼/2𝑛, n-p-1) = 2.246

Two-sided Interval : bk ± B\*s{bk}

b1= 10.99 b2=0.566 s{b1}=3.56 s{b1}=0.214

C.I for Palm Width: ( 2.994, 18.98576)

C.I for Height: (0.0853, 1.046644)

We will now obtain 95% CI , PI ,CB for Palm width of 9.5 and height of 170.

|  |  |  |  |
| --- | --- | --- | --- |
| CLIM | CLIM\_1 | PLIM | PLIM\_1 |
| 45.2985 | 69.97959 | 36.26737 | 79.01071 |

# Conclusion:

In this regression model we started with 4 predictor variables palm width, Palm Length , Height, Weight. We Performed all the preliminary analysis to Find out if the model was adequate. There was a slight curvture for the weight plot so we dwcided to consider weight2 predictor variable. We experimented with Interaction terms by multiplying predictor variables among themselves to find out if they would result in a better regression model. We found out that the corelation was high. So we considered using standarized predictor variable. We searched for the best model using backward deletion, stepwise regression and best subsets method. By doing this we narrowed the different regreesion model to 2 model:

Model 1 : Grip strength (y) = -143.0 + 10.99 Palm Width + 0.566 Height

Model 2 : Model2: Grip strength = -158.38 + 6.81 Palm width + 2.17Weightsq + 8.95 STD X1X3 – 28.37 STD X4X1

Then Among this too model we selected the best model with reference to above mentioned criterias.

The best model was:

Model 1: Grip strength (y) = -143.0 + 10.99 Palm Width + 0.566 Height

We did all the necessary tests and concluded that model1 was the best model.

# APPENDIX

## BACKWARD REGRESSION

Start: AIC=139.67

Grip.Strength ~ Palm.Length + Palm.Width + Height + Weight^2 +

stdx1x2 + stdx4x1 + stdx2x4 + stdx1x3

Df Sum of Sq RSS AIC

- Palm.Length 1 0.001 1731.8 137.67

- stdx1x3 1 0.188 1732.0 137.68

- Palm.Width 1 0.209 1732.0 137.68

- Height 1 0.545 1732.4 137.68

- stdx1x2 1 0.955 1732.8 137.69

- stdx2x4 1 3.803 1735.6 137.74

- stdx4x1 1 16.642 1748.5 137.96

- Weight 1 43.889 1775.7 138.42

<none> 1731.8 139.67

Step: AIC=137.67

Grip.Strength ~ Palm.Width + Height + Weight + stdx1x2 + stdx4x1 +

stdx2x4 + stdx1x3

Df Sum of Sq RSS AIC

- Palm.Width 1 0.215 1732.0 135.68

- stdx1x3 1 0.350 1732.2 135.68

- Height 1 0.862 1732.7 135.69

- stdx1x2 1 1.322 1733.2 135.69

- stdx2x4 1 7.241 1739.1 135.80

- stdx4x1 1 17.026 1748.9 135.97

- Weight 1 55.352 1787.2 136.62

<none> 1731.8 137.67

Step: AIC=135.68

Grip.Strength ~ Height + Weight + stdx1x2 + stdx4x1 + stdx2x4 +

stdx1x3

Df Sum of Sq RSS AIC

- stdx1x3 1 0.145 1732.2 133.68

- Height 1 2.369 1734.4 133.72

- stdx2x4 1 7.157 1739.2 133.80

- stdx1x2 1 21.701 1753.8 134.05

- stdx4x1 1 43.028 1775.1 134.41

- Weight 1 98.577 1830.6 135.34

<none> 1732.0 135.68

Step: AIC=133.68

Grip.Strength ~ Height + Weight + stdx1x2 + stdx4x1 + stdx2x4

Df Sum of Sq RSS AIC

- stdx2x4 1 26.26 1758.5 132.13

- Weight 1 99.96 1832.2 133.36

- stdx1x2 1 106.09 1838.3 133.46

<none> 1732.2 133.68

- stdx4x1 1 165.81 1898.0 134.42

- Height 1 362.42 2094.6 137.38

Step: AIC=132.13

Grip.Strength ~ Height + Weight + stdx1x2 + stdx4x1

Df Sum of Sq RSS AIC

<none> 1758.5 132.13

- stdx1x2 1 239.96 1998.4 133.97

- stdx4x1 1 267.98 2026.4 134.38

- Weight 1 369.64 2128.1 135.85

- Height 1 416.79 2175.2 136.51

Call:

lm(formula = Grip.Strength ~ Height + Weight + stdx1x2 + stdx4x1,

data = data)

Coefficients:

(Intercept) Height Weight stdx1x2 stdx4x1

-170.3229 0.5455 1.7698 7.0051 -22.4126