1. Introduction

Computers are of a mere need in day to day life and any everyday there are new advancements in computers. The hardware of computers has a vital role in the evolution of computers by improving its process ability and capability. Measuring the performance of hardware helps the engineers and scientists to build better hardwares and to innovate new potential. To enhance the same, a regression model is developed in RStudio to study the effect of seven input variables (machine cycle time, minimum main memory, maximum main memory, cache memory, minimum channels, maximum channels, published relative performance) on output variable estimated relative performance. The relationship of the ingredients and the strength is estimated with the help of regression analysis.

2. Purpose

The primary intention of the project is to find all the influential parameters for getting the best response variable. Thorough data analysis must be conducted by checking for **data transformation, model adequacy, multicollinearity, Test of significance, variable selection and model validation** on the initial model**.** The Final model should be able to provide good fit with predicting capability of all the future observations.

3. Dataset

The data is obtained from "UCI machine learning repository". The dataset was created by Phillip Ein-Dor and Jacob Feldmesser (Ein-Dor: Faculty of Management, Tel Aviv University; Ramat-Aviv; Tel Aviv, 69978; Israel) and was donated by David W. Aha (aha **'@'** ics.uci.edu (714) 856-8779). In total the dataset consists of 209 observations in which 7 variables (denoted by X1…X7) aims to predict output variable (denoted by Y). Following is the attribute information:

Specifically:

Regressors/Response	Information
X1	Machine cycle time
X2	Minimum main memory
X3	Maximum main memory
X4	Cache memory
X5	Minimum channels
X6	Maximum channels

X7	Published relative performance(PRF)
Y	Estimated relative performance(ERF)

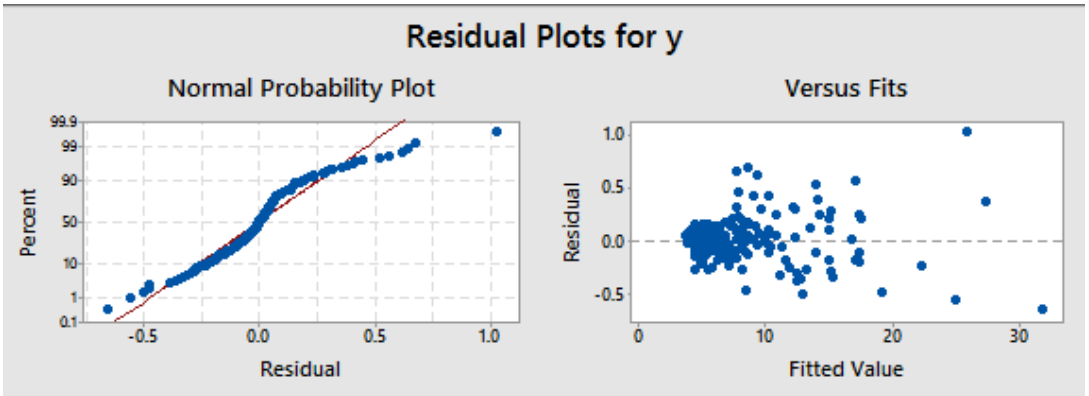
4. Data Analysis

A linear model is generated using all the predictor variables. This model is analysed for P value and adjusted R square using RStudio and Minitab.

4.1 Initial model:

```
> hardproject <- read.csv(file.choose(),header=T)
> modelhardproject <- lm(y~.,hardproject)
> summary(modelhardproject)

Call:
lm(formula = y ~ ., data = hardproject)

Residuals:
     Min       1Q   Median       3Q      Max
-117.478   -9.546    2.864   15.257  182.251

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.423e+01  4.732e+00  -7.234 9.68e-12 ***
x1           3.777e-02  9.434e-03   4.004 8.77e-05 ***
x2           5.483e-03  1.120e-03   4.894 2.02e-06 ***
x3           3.375e-03  3.974e-04   8.493 4.45e-15 ***
x4           1.244e-01  7.751e-02   1.605  0.11016
x5          -1.634e-02  4.523e-01  -0.036  0.97122
x6           3.458e-01  1.287e-01   2.687  0.00781 **
x7           5.770e-01  3.718e-02  15.519  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 31.7 on 201 degrees of freedom
Multiple R-squared:  0.9595,    Adjusted R-squared:  0.958
F-statistic: 679.5 on 7 and 201 DF,  p-value: < 2.2e-16

> print(vif(modelhardproject))
      x1       x2       x3       x4       x5       x6       x7
1.247906 3.908747 4.495366 2.052662 1.967205 2.316409 7.401415
```

Table. Model Summary of Initial Model



Result: The Normal probability plot doesn't seem to be perfect as the **values don't fall along the straight line** and there are few outliers forming a tail. There is a non-constant variance as there is a **funnel opening** facing outside in residual vs. fitted value plot. Suitable data transformation should be applied in this case. R squared and Adjusted R squared values are 95.95% and 95.8% respectively which shows the model does not effectively explain the variability in the response. The table states that the Cache memory and Minimum channels are not significant predictors in this analysis. There is also a moderate multicollinearity present in the model with **maximum VIF of 7.4014**.

4.2 Re-Specified Model: Model re-specification is done based by removing the multicollinearity on the model. The x7 regressor has a vif value of 7.4104 and it is divided by the next largest regressor x3 to remove the multicollinearity in the model **(xnew=(x7/x3))**. The scatter plots suggest that the x1 regressor (machine cycle time) isn't following a linear relationship with the response y. We transform the regressor **x1 to (1/x1).**

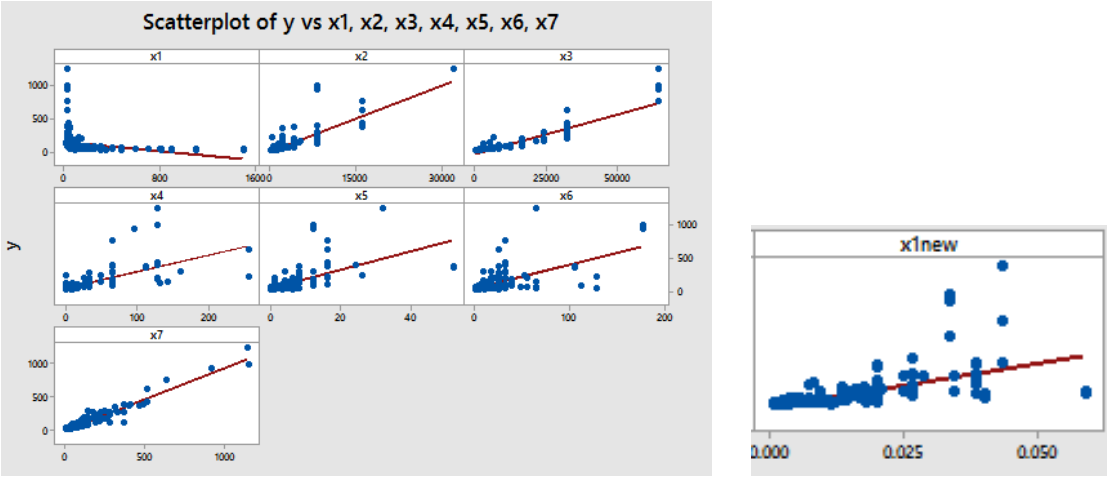


Fig. Scatter Plot of Regressors vs Response and corrected Scatter plot of x1 vs y

Box Cox method is employed to understand the suitable data transformation and the transformation value of the response y is found to be around 0.5 **(ynew=sqrt(y))**.

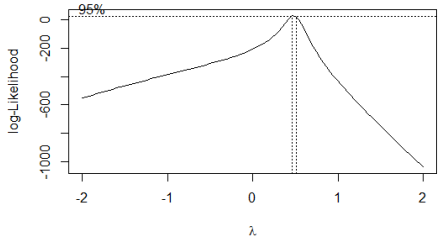


Fig. Box-Cox transformation

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.064e+00  1.901e-01  21.375  < 2e-16 ***
x1new        7.424e+01  1.365e+01   5.437 1.55e-07 ***
x2           6.438e-04  4.505e-05  14.291  < 2e-16 ***
x4           3.183e-02  4.018e-03   7.923 1.53e-13 ***
x5           2.731e-02  2.391e-02   1.142 0.254675
x6           5.355e-02  5.514e-03   9.711  < 2e-16 ***
xnew        -2.596e+01  7.502e+00  -3.460 0.000659 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.657 on 202 degrees of freedom
Multiple R-squared:  0.8981,    Adjusted R-squared:  0.8951
F-statistic: 296.7 on 6 and 202 DF,  p-value: < 2.2e-16
```

```
> print(vif(modelhardwarefinal))
   x1new       x2       x4       x6     xnew
2.034166 2.238138 1.914412 1.315872 1.125832
```

Table. Model Summary and VIF's of Re-Specified model

Result: The R-squared and Adjusted R-squared are improved from values ranging 83 to 89.91 and 89.51 respectively as compared to the model without transformation and with multicollinearity removed. The plot of Residuals vs. fitted value is much better than the previous model **overcoming the non-constant variance problem. The** Multicollinearity problem has been solved by reducing the VIF's less than 5 in the model. The plot of residuals vs. Log(X8) shows a better variance than the initial plot. Plot of residuals vs. rest of the variables were also checked and there was no problem detected.

4.3 Variable Selection and Model Building:

```
Response is ynew

                                                    x
                                                  x 1
                                                  n n
             R-Sq            R-Sq  Mallows         e e x x x x
Vars  R-Sq  (adj)   PRESS  (pred)       Cp       S  w w 6 5 4 2
   1  70.0   69.9  1697.7    68.8    388.8  2.8072             X
   1  53.5   53.3  2733.7    49.8    716.7  3.4973           X
   2  83.8   83.7   982.6    82.0    117.6  2.0677       X     X
   2  81.3   81.1  1100.6    79.8    167.0  2.2214           X X
   3  87.7   87.5   794.9    85.4     42.5  1.8064       X   X X
   3  86.1   85.9   838.3    84.6     74.0  1.9196     X X     X
   4  89.2   88.9   697.2    87.2     16.0  1.7017     X X   X X
   4  88.1   87.9   790.7    85.5     36.7  1.7814   X   X   X X
   5  89.7   89.5   690.8    87.3      6.3  1.6587   X X X   X X
   5  89.2   88.9   718.2    86.8     17.0  1.7016     X X X X X
   6  89.8   89.5   709.2    87.0      7.0  1.6574   X X X X X X
```

Best Subsets Regression: sqrt(y) versus x1new=(1/x),x2, x4, x5, x6, xnew=(x3/x7)

Result: The highlighted rows shows two best subset models that can be considered. Since the Adjusted R-square seems to be constant both the subsets. Finally the subset model with **low Mallows Cp 6.3** is being removed and the obtained model is considered as the **final model with five regressors namely x1new=(1/x)(Machine cycle time), x2(Minimum main memory),**

X4(Cache memory), x6(Maximum channels) and xnew=(x7/x3) (PRF/Max main memory). This final model is subjected to thorough analysis and its details using R is shown below.

```
Call:
lm(formula = ynew ~ x1new + x2 + x4 + x6 + xnew)

Residuals:
    Min      1Q  Median      3Q     Max
-6.1380 -0.9398 -0.0659  0.8275  5.8849

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.058e+00  1.902e-01  21.336  < 2e-16 ***
x1new        7.670e+01  1.349e+01   5.685 4.52e-08 ***
x2           6.531e-04  4.436e-05  14.722  < 2e-16 ***
x4           3.287e-02  3.917e-03   8.392 8.10e-15 ***
x6           5.603e-02  5.075e-03  11.040  < 2e-16 ***
xnew        -2.568e+01  7.503e+00  -3.422 0.000751 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.659 on 203 degrees of freedom
Multiple R-squared:  0.8974,    Adjusted R-squared:  0.8949
F-statistic: 355.2 on 5 and 203 DF,  p-value: < 2.2e-16
```

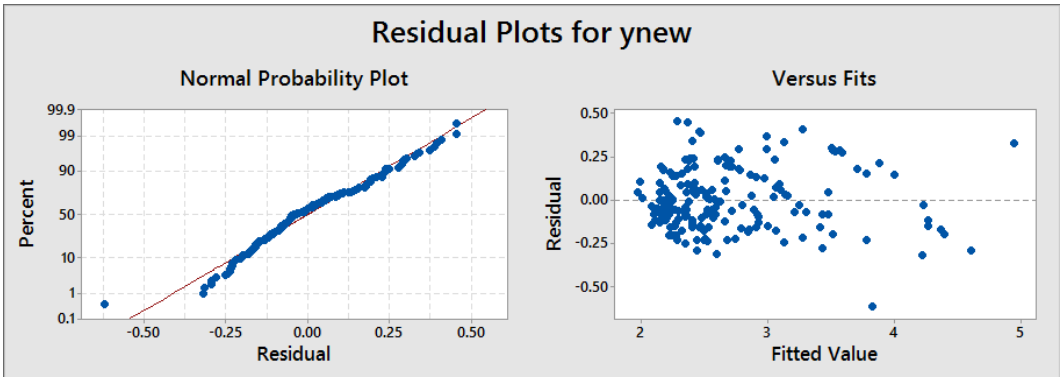


Fig. Final Model Summary and Model adequacy

Result: The Summary shows all five regressors being significant now. The R square and R square adjusted values remains constant and not much change is seen after performing variable selection and model building. The normal probability plot has been updated and a linear normal plot can be seen. The Residual vs fit does not have any shape or pattern to carry. The normality plot has presence of some influential and leverage point and we can't simply remove these points as they are valid observations. Further the Model is validated and the stability of the co-efficients are calculated.

5. Conclusion and Future Scope:

The simple linear equation was fit for the data and it showed abnormality. Then the Multicolimearity was eliminated by model re-specification. The VIF's were reduced below

five and confirmed the same. The data was then transformed using suitable techniques and analysis of transformed data yielded R-squared of 89.74% and Adjusted R-squared of 89.49% which is a fair result. All possible techniques were used to remove Insignificant variables in the model. The sign and magnitude of variable coefficients validated the final model and thus the model can be recommended to any computer hardware developer. The above analysis has a limitation in the value of Adjusted R squared to 89.49% even after all the iterations of removal of insignificant variables. In Future, **Principal Component Regression** analysis will be carried out to estimate if total variance explained by the model will be better than this analysis or not.

6.Bibliography:

1. Montgomery, D. C., Peck Elizabeth, A., & Geoffrey Vining, G. Introduction to Linear Regression Analysis, Fifth Edition.

2. http://rexa.info/paper/3caf773de7b1ad6c9236cbd03763058bc8846e9d

3. https://archive.ics.uci.edu/ml/datasets/Computer+Hardware

4. https://www.lifewire.com/computer-hardware-2625895