VIETNAM NATIONAL UNIVERSITY, HO CHI MINH CITY
HO CHI MINH CITY UNIVERSITY OF TECHNOLOGY
Faculty of Computer Science and Engineering

PROBABILITY AND STATISTICS (MT 2013)

Assignment

"CPU relative performance "

Advisor: Dr. Phan Thi Huong
Students: Pham Duc Trung - 2153928
Nguyen Phuoc Thinh - 2153838
Tran Hai Dang - 2153297
Le Duong Khanh Huy - 2153380
Phan Le Khanh Trinh - 2151268
Class: CC01 - Group : 7

HO CHI MINH CITY, April 2023

Group contribution

No.	**Fullname**	**Student ID**	**Percentage of work**
1	Pham Duc Trung	2153928	20%
2	Nguyen Phuoc Thinh	2153838	20%
3	Tran Hai Dang	2153297	20%
4	Le Duong Khanh Huy	2153380	20%
5	Phan Le Khanh Trinh	2151268	20%

Leader: Nguyen Phuoc Thinh - thinh.nguyenph@hcmut.edu.vn

Lecturer's assessment

No.	**Section**	**Student ID**	**Evaluation**
1	Pham Duc Trung	2153928	
2	Nguyen Phuoc Thinh	2153838	
3	Tran Hai Dang	2153297	
4	Le Duong Khanh Huy	2153380	
5	Phan Le Khanh Trinh	2151268	
6	Total		

Contents

1 Data introduction

This assignment presents an analysis of computer system performance, focusing on the relationship between machine characteristics and relative performance metrics. The dataset includes attributes:

1. **Vendor Name:** many unique symbols
2. **Model Name:** many unique symbols
3. **MYCT:** machine cycle time in nanoseconds (integer)
4. **MMIN:** minimum main memory in kilobytes (integer)
5. **MMAX:** maximum main memory in kilobytes (integer)
6. **CACH:** cache memory in kilobytes (integer)
7. **CHMIN:** minimum channels in units (integer)
8. **CHMAX:** maximum channels in units (integer)
9. **PRP:** published relative performance (integer)
10. **ERP:** estimated relative performance from the original article (integer)

Listing 1: Data set factors

In this assignment, we totally use R and R Studio as tools and working environment to analyze data.

	vendor_name	model_name	MYCT	MMIN	MMAX	CACH	CHMIN	CHMAX	PRP	ERP
1	adviser	32/60	125	256	6000	256	16	128	198	199
2	amdahl	470v/7	29	8000	32000	32	8	32	269	253
3	amdahl	470v/7a	29	8000	32000	32	8	32	220	253
4	amdahl	470v/7b	29	8000	32000	32	8	32	172	253

Figure 1: Piece of Labeled Data Frame

Check out some key factors in this data set:
population : CPU relative performance
sample : 209 models CPU relative performance measured by dataset creator.
parameter : MYCT, MMIN, MMAX, CACH, CHMIN, CHMAX, PRP ,ERP.
categorical variables : vendor_name, model_name.

In this data set, we spend much attention on two variables PRP and ERP. At first glance, there are many parameters related to each model seem to have impacts on ERP and PRP (which are model performance), so regression models would be our methods to approach the data set in this assignment.

2 Background

2.1 Exploratory data analysis (EDA)

Exploratory data analysis (EDA) is a method for drawing insights from data, often utilizing data visualization and statistical graphics to reveal relationships between variables, identify patterns and trends, and detect outliers. EDA is crucial for extracting important features for predictive models. By plotting the raw data, we can gain an understanding of the general behavior and distribution of the variables:

- Histograms are used to visualize the distribution of a numerical variable.
- Box plots are used to display the distributions of numerical data values, particularly when comparing them across several groups.
- Pair plots are employed to comprehend the finest characteristics that may be applied to describe a link between two variables or to create the most distinct clusters.
- Correlation Matrix: The correlation coefficients between variables are displayed in a table called a correlation matrix. The association between two variables is displayed in each cell of the table.

2.2 Multivariate Linear Regression (MLR)

There isn't always a response variable or an explanatory variable in statistical analyses of data. Measurement of the association between a continuous dependent variable and two or more independent variables is done using the multivariate regression method. Linear relationships are those that emerge from correlations between variables. We utilize this method to forecast the behavior of a response variable based on its predictors after applying multivariate regression to a data set.

There are 3 assumptions that need to be met when performing an MLR test:

1. Normality: the residuals (the differences between the observed values and the predicted values) follow a normal distribution.
2. No multicollinearity: The independent variables in the regression model are not highly correlated with each other. Multicollinearity occurs when two or more independent variables are strongly correlated, making it difficult to determine their individual effects on the dependent variable.
3. Linearity : The relationship between the independent variables and the dependent variable should be linear. This means that the expected value of the dependent variable changes in a straight line as the independent variables change, holding other variables constant.

The general equation of MLR:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n + \epsilon$$

where :

·Y is the dependent variable.
·X_i is the i_{th} independent variable.
·B_0 is the intercept of Y when all X_i are zeroes.
·B_i is the coefficient of each X_i.
·ϵ is the independent error term for the model.

Model performance metrics :

· R-squared (R^2): is the squared correlation between the actual result value and the value predicted by the model, and it measures the proportion of the predictor's variance in the outcome that can be accounted for. The better the model, the higher the value.

· Root mean square error (RMSE): calculates the average error a model makes while forecasting an observation. The model is better the lower the RMSE.

3 Descriptive statistic

3.1 Data pre-processing

• Importing data

Importing file name "machine.data.txt" into data frame (noted as **df** in R).

	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10
1	adviser	32/60	125	256	6000	256	16	128	198	199
2	amdahl	470v/7	29	8000	32000	32	8	32	269	253
3	amdahl	470v/7a	29	8000	32000	32	8	32	220	253
4	amdahl	470v/7b	29	8000	32000	32	8	32	172	253

Figure 2: Initial Data Frame

• Naming data feature

As we can see , columns in this data frame are not labeled. Let's make the data frame follow the list we have above (**Listing 1**).

	vendor_name	model_name	MYCT	MMIN	MMAX	CACH	CHMIN	CHMAX	PRP	ERP
1	adviser	32/60	125	256	6000	256	16	128	198	199
2	amdahl	470v/7	29	8000	32000	32	8	32	269	253
3	amdahl	470v/7a	29	8000	32000	32	8	32	220	253
4	amdahl	470v/7b	29	8000	32000	32	8	32	172	253

Figure 3: Labeled Data Frame

• Checking missing value

Now it seems find. However, to make sure there is no problem in this data set, we implement some function to test whether they are any missing data or relative problems.

Result

```
vendor_name  model_name        MYCT        MMIN        MMAX        CACH
          0           0           0           0           0           0
      CHMIN       CHMAX         PRP         ERP
          0           0           0           0
```

Figure 4: Checking missing value

As can be seen from figure above there is no N/A value in our dataset, and it is suitable for the next step.

3.2 Data summary

After having done cleaning process, we currently have a clear and clean data set in the data frame(**Figure 2**). Let's summary **df** by using function summary in R.

Result

```
vendor_name          model_name              MYCT             MMIN
ibm       : 32       100        :  1    Min.   :  17.0   Min.   :   64
nas       : 19       1100/61-h1:  1    1st Qu.:  50.0   1st Qu.:  768
honeywell: 13       1100/81    :  1    Median : 110.0   Median : 2000
ncr       : 13       1100/82    :  1    Mean   : 203.8   Mean   : 2868
sperry    : 13       1100/83    :  1    3rd Qu.: 225.0   3rd Qu.: 4000
siemens   : 12       1100/84    :  1    Max.   :1500.0   Max.   :32000
(Other)   :107       (Other)    :203
      MMAX             CACH             CHMIN            CHMAX
Min.   :   64      Min.   :  0.00     Min.   : 0.000   Min.   :  0.00
1st Qu.: 4000      1st Qu.:  0.00     1st Qu.: 1.000   1st Qu.:  5.00
Median : 8000      Median :  8.00     Median : 2.000   Median :  8.00
Mean   :11796      Mean   : 25.21     Mean   : 4.699   Mean   : 18.27
3rd Qu.:16000      3rd Qu.: 32.00     3rd Qu.: 6.000   3rd Qu.: 24.00
Max.   :64000      Max.   :256.00     Max.   :52.000   Max.   :176.00
      PRP               ERP
Min.   :   6.0     Min.   :  15.00
1st Qu.:  27.0     1st Qu.:  28.00
Median :  50.0     Median :  45.00
Mean   : 105.6     Mean   :  99.33
3rd Qu.: 113.0     3rd Qu.: 101.00
Max.   :1150.0     Max.   :1238.00
```

Figure 5: Data overview

Because Vendor Name and Model Name are categorical variables, we can not have an overview information in this variable, so we try to make it clearer.

Result

```
Number of vendors : 30
Most number of models : ibm :  32
Least number of model(s):  adviser, four-phase, microdata, sratus :  1
```

Figure 6: Factorial summary

We take a clearer look at the distribution of the 8 independent variables, utilizing the Histogram and Boxplot and applying the function to each figure variable.

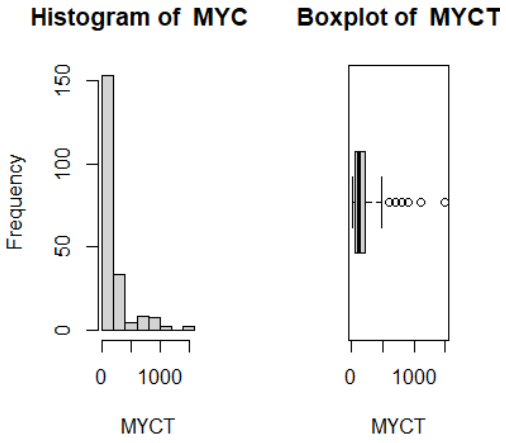

Figure 7: MYCT

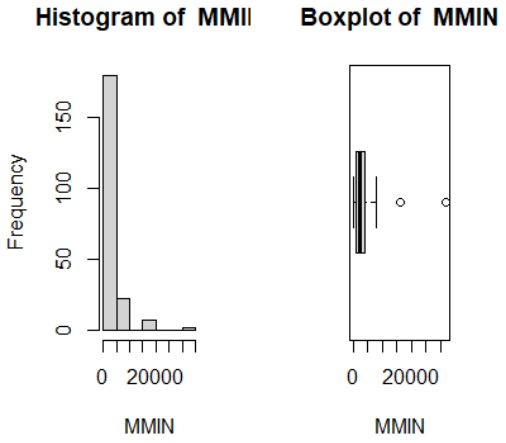

Figure 8: MMIN

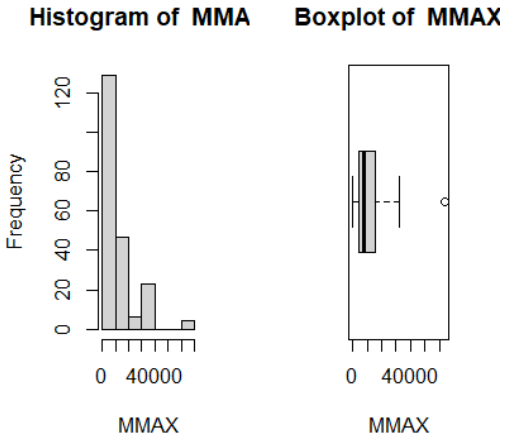

Figure 9: MMAX

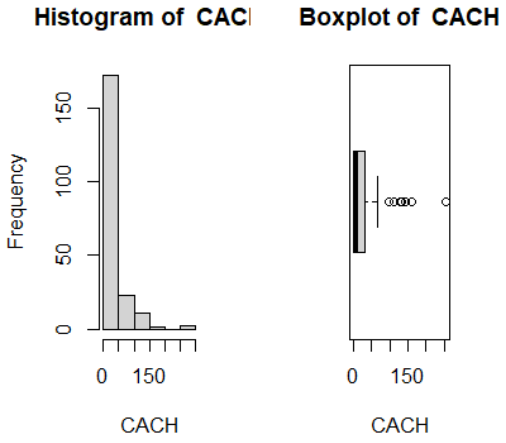

Figure 10: CACH

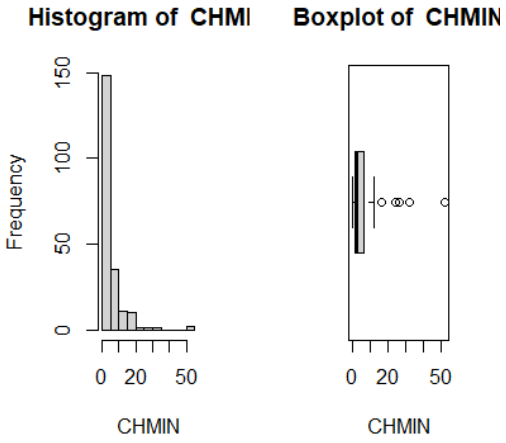

Figure 11: CHMIN

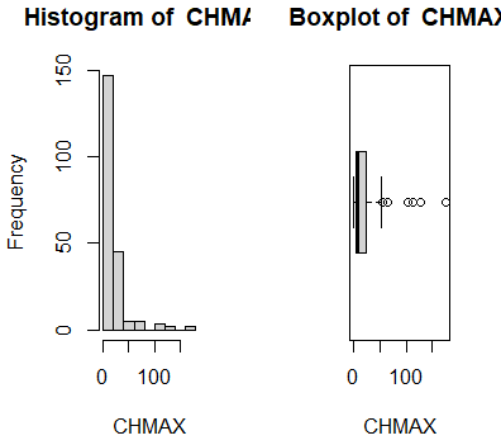

Figure 12: CHMAX

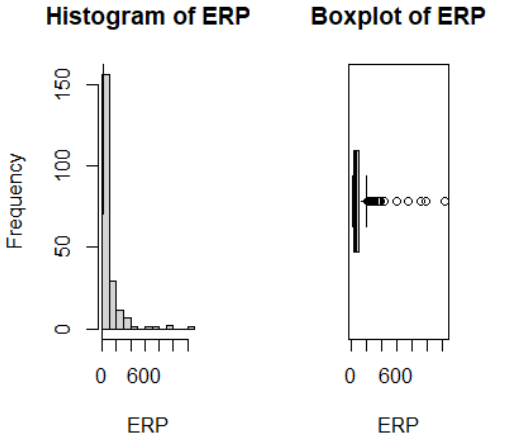

Figure 13: PRP

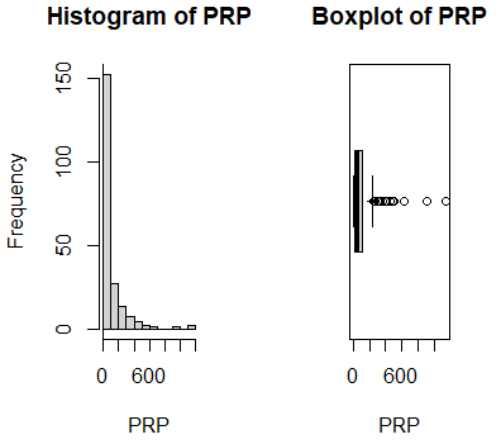

Figure 14: ERP

According to the summary and these plots above:

- **vendor_name**: The dataset contains several different vendors. The most common vendor is IBM, with 32 machines, followed by NAS with 19 machines. The 'Other' category contains 107 machines from various different vendors.
- **model_name**: There are a variety of model names in the dataset, with no model appearing more than once except for the models under the 'Other' category. This indicates a diverse range of models in the dataset.
- **MYCT**: Machine cycle time (MYCT) ranges from 17 to 1500 with a median of 110. As the mean (203.8) is larger than the median, this distribution is right-skewed, suggesting there are some machines with particularly high cycle times.

- **MMIN**: Minimum main memory (MMIN) ranges from 64 to 32000. The mean (2868) is greater than the median (2000), indicating a right-skewed distribution. This suggests that some machines have particularly high minimum memory.
- **MMAX**: Maximum main memory (MMAX) ranges from 64 to 64000. Like MMIN, the mean (11796) is greater than the median (8000), indicating a right-skewed distribution. Some machines have exceptionally high maximum memory.
- **CACH**: Cache memory size (CACH) ranges from 0 to 256. The mean (25.21) is larger than the median (8), so this distribution is right-skewed. Some machines have very large cache memory sizes.
- **CHMIN and CHMAX**: The minimum and maximum numbers of channels (CHMIN, CHMAX) also have right-skewed distributions, with means larger than their medians. This suggests that some machines have an unusually high number of channels.
- **PRP and ERP**: The published and estimated relative performance measures (PRP, ERP) both have right-skewed distributions, with means larger than the medians. This suggests some machines have particularly high performance.

Then, we use Correlation plot to give the overview of the relation between each pair of variable.

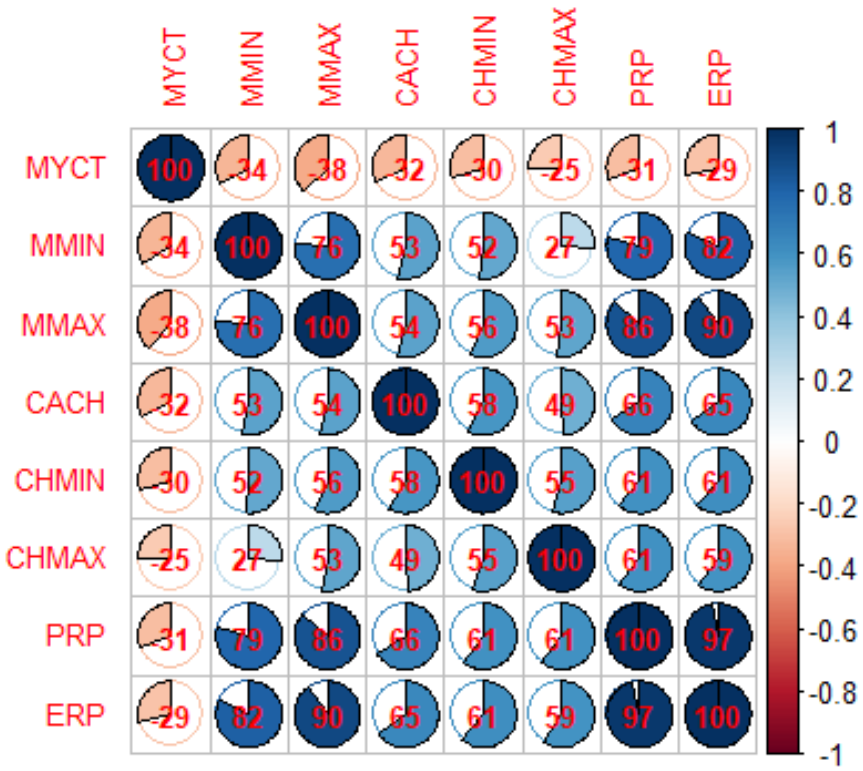

Figure 15: Correlation Plot

According to the correlation plot above, we can see:

- **MYCT**: This variable is negatively correlated with all other variables, with the strongest negative correlations with MMIN, MMAX, and CACH. This suggests that as the machine cycle time (MYCT) increases, these other measures tend to decrease.
- **MMIN**: There is a very strong positive correlation between MMIN (minimum main memory) and ERP (Estimated Relative Performance), MMAX (maximum main memory), and PRP (Published Relative Performance), which suggests that machines with more minimum main memory tend to have higher performance and more maximum memory.
- **MMAX**: This variable has a strong positive correlation with ERP and PRP, indicating that machines with more maximum memory tend to have better performance. There is also a strong positive correlation with MMIN, suggesting that machines often have similar amounts of minimum and maximum memory.
- **CACH**: Cache size (CACH) has strong positive correlations with PRP, ERP, MMIN, and MMAX, suggesting that machines with more cache tend to have more memory and better performance.
- **CHMIN**: The minimum channels (CHMIN) variable is moderately positively correlated with all other variables except MYCT. The strongest correlations are with ERP, PRP, and MMAX, indicating that machines with more minimum channels tend to have more maximum memory and better performance.
- **CHMAX**: Maximum channels (CHMAX) is positively correlated with all other variables except MYCT. The strongest correlations are with PRP, ERP, and MMIN, suggesting that machines with more maximum channels tend to have more minimum memory and better performance.
- **PRP and ERP**: These two measures of performance are extremely strongly correlated with each other, suggesting they are measuring largely the same construct. They are also both strongly positively correlated with MMIN and MMAX, indicating that machines with more memory tend to have better performance.

Many of the variables appear to have strong associations with both PRP and ERP. Going forward, we will investigate techniques for analyzing the connections between these variables and PRP.

4 Inferential statistic

4.1 Assumption

4.1.1 Normality

We will choose Shapiro-Wilk test to ensure residual is normal distributed.

Result

```
    Shapiro-Wilk normality test
    data:  residuals
    W = 0.83127, p-value = 2.597e-14
```

Figure 16: Shapiro-Wilk test result

The p value of residual is approximately equal zero. So, we can make certain that the residual after using linear regression model is normally distributed.

4.1.2 Linearity

Result

```
    Call:
    lm(formula = PRP ~ MYCT + MMIN + MMAX + CACH + CHMIN + CHMAX,
    data = df)
    Residuals:
    Min      1Q      Median      3Q      Max
    -195.82  -25.17    5.40       26.52    385.75

Coefficients:
              Estimate Std. Error t value Pr(>t)
(Intercept) -5.589e+01  8.045e+00  -6.948 5.00e-11 ***
MYCT         4.885e-02  1.752e-02   2.789   0.0058 **
MMIN         1.529e-02  1.827e-03   8.371 9.42e-15 ***
MMAX         5.571e-03  6.418e-04   8.681 1.32e-15 ***
CACH         6.414e-01  1.396e-01   4.596 7.59e-06 ***
CHMIN       -2.704e-01  8.557e-01  -0.316   0.7524
CHMAX        1.482e+00  2.200e-01   6.737 1.65e-10 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1  1

Residual standard error: 59.99 on 202 degrees of freedom
Multiple R-squared:  0.8649,    Adjusted R-squared:  0.8609
F-statistic: 215.5 on 6 and 202 DF,  p-value: < 2.2e-16
```

Figure 17: Check for Linearity

In linear regression, the coefficient of determination, often referred to as R-squared (R^2), is a statistical measure that represents the proportion of the variance in the dependent variable that can be explained by the independent variables in the model. R-squared ranges from 0 to 1
In this project, R-squared = 0.8649 (near 1), which means a large proportion of the variability in the dependent variable can be explained by the independent variables included in the model. So we can conclude that there is a linear relationship between PRP and others variables.

4.1.3 Multicollinearity check

Result

```
     MYCT      MMIN      MMAX          CACH      CHMAX
 1.199030  2.792328  3.266378      1.730247  1.691608
```

Figure 18: Coefficient of variation

Multicollinearity has been checked for the variables in the model using the Variance Inflation Factor (VIF). The VIF is a measure of how much the variance of the estimated regression coefficients are increased due to multicollinearity.

The VIF for the variables are as follows: MYCT = 1.199, MMIN = 2.792, MMAX = 3.266, CACH = 1.730, and CHMAX = 1.692.

Generally, a VIF greater than 5 or 10 indicates a problematic amount of multicollinearity. In this case, none of the variables have a VIF above 5, which suggests that multicollinearity is likely not a problem in this model.

4.2 Multivariate Linear Regression (MLR)

In this section, we will build a multiple linear regression model with PRP as the dependent variable, and MYCT, MMIN, MMAX, CACH, CHMIN, and CHMAX as the independent variables. Our objective is to investigate the relationship between PRP and these independent variables, and to develop a predictive model that accurately estimates PRP based on these factors. This analysis will provide valuable insights into the factors that significantly impact computer system performance and facilitate optimization of system configurations in the future.

We fitted our statistical model using the lm function in R. The lm function allows us to estimate the relationship between a dependent variable and one or more independent variables.

4.2.1 Splitting data

In machine learning, splitting the data into distinct sets helps prevent overfitting, which occurs when a model memorizes the training data too well and fails to generalize to new, unseen data. So that, we split the data into training and testing sets by 80% and 20% amount of data.

4.2.2 Fitting the multivariate linear regression model

Result

```
    Residuals:
    Min      1Q      Median      3Q      Max
    -195.82  -25.17   5.40       26.52    385.75

Coefficients:
              Estimate Std. Error t value Pr(>t)
(Intercept) -5.589e+01  8.045e+00  -6.948 5.00e-11 ***
MYCT         4.885e-02  1.752e-02   2.789   0.0058 **
MMIN         1.529e-02  1.827e-03   8.371 9.42e-15 ***
MMAX         5.571e-03  6.418e-04   8.681 1.32e-15 ***
CACH         6.414e-01  1.396e-01   4.596 7.59e-06 ***
CHMIN       -2.704e-01  8.557e-01  -0.316   0.7524
CHMAX        1.482e+00  2.200e-01   6.737 1.65e-10 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1  1

Residual standard error: 59.99 on 202 degrees of freedom
Multiple R-squared:  0.8649,    Adjusted R-squared:  0.8609
F-statistic: 215.5 on 6 and 202 DF,  p-value: < 2.2e-16
```

Figure 19: Multivariate linear regression model - Result

As can be seen from the result:

- The multivariate linear regression model was fitted, and the residuals range from approximately -195.82 to 385.75, with the 1st quartile, median, and 3rd quartile being 25.17, 5.40, and 26.52, respectively. This suggests that there is considerable spread in the prediction errors, including some particularly large positive and negative residuals.
- The estimated coefficients (or parameters) of the model show the direction and magnitude of the relationship between the predictor variables (MYCT, MMIN, MMAX, CACH, CHMIN, CHMAX) and the response variable. In this case, all the predictors have positive relationships with the response except for CHMIN, which is negative. However, the CHMIN predictor is not significant (p-value = 0.7524) as it is greater than a typical level of 0.05. This suggests that CHMIN does not significantly contribute to the model.
- The remaining variables (MYCT, MMIN, MMAX, CACH, CHMAX) have p-values below 0.05, indicating that they are statistically significant predictors of the response. The smallest p-value is observed for CHMAX, indicating that it is a highly significant predictor.
- The model's Residual Standard Error (RSE) is approximately 59.99. This value is a measure of the typical deviation of the actual values from the predicted values.
- The multiple R-squared (0.8649) and the Adjusted R-squared (0.8609) values suggest that the model explains a significant proportion of the variability in the response. The adjusted R-squared, which penalizes the addition of unnecessary predictors in the model, is slightly lower than the multiple R-squared, indicating a good model fit.
- The overall F-statistic (215.5) and its corresponding p-value (less than 2.2e-16) test the hypothesis that at least one of the predictors is useful in explaining the response. Given the p-value is very small, we reject the null hypothesis, indicating that at least one predictor is contributing significantly to the prediction of the response.

4.2.3 Stepwise Regression

Next, we will apply Stepwise regression to choose the most suitable model.

Result

```
    Start:  AIC=1718.24
PRP ~ MYCT + MMIN + MMAX + CACH + CHMIN + CHMAX

         Df Sum of Sq    RSS    AIC
- CHMIN   1       359 727279 1716.3
<none>                726920 1718.2
- MYCT    1     27985 754905 1724.1
- CACH    1     76009 802929 1737.0
- CHMAX   1    163347 890267 1758.6
- MMIN    1    252179 979099 1778.5
- MMAX    1    271177 998097 1782.5
Multiple R-squared:  0.8649,    Adjusted R-squared:  0.8609

Step:  AIC=1716.34
PRP ~ MYCT + MMIN + MMAX + CACH + CHMAX

         Df Sum of Sq    RSS    AIC
<none>                727279 1716.3
+ CHMIN   1       359 726920 1718.2
- MYCT    1     28343 755623 1722.3
- CACH    1     78715 805995 1735.8
- CHMAX   1    177114 904393 1759.9
- MMIN    1    258252 985531 1777.8
- MMAX    1    270856 998135 1780.5
Multiple R-squared:  0.8648,    Adjusted R-squared:  0.8615
```

Figure 20: Stepwise regression result

As we can see from the result above:

- The stepwise regression model fitting procedure has been applied to the initial linear regression model (model1) which included all variables (MYCT, MMIN, MMAX, CACH, CHMIN, CHMAX). The Akaike Information Criterion (AIC) has been used as the measure to compare different models. A lower AIC value indicates a better fitting model.
- The AIC of the initial model was 1718.24. Stepwise regression procedure started with this full model and considered the effect of removing each variable on the AIC.
- The removal of variable CHMIN would lead to the smallest increase in the Residual Sum of Squares (RSS) and would also reduce the AIC from 1718.24 to 1716.3. Therefore, in the first step, CHMIN was removed from the model.
- The stepwise procedure continued by considering adding the variable CHMIN back into the model or removing any of the remaining variables (MYCT, MMIN, MMAX, CACH, CHMAX). However, adding CHMIN back would increase the AIC, and removing any of the remaining variables would also increase the AIC.

Therefore, the procedure stopped, and the final model included MYCT, MMIN, MMAX, CACH, and CHMAX but not CHMIN. The AIC of the final model was 1716.3, which is lower than the AIC of the initial model, indicating that the stepwise-selected model provides a better fit

according to this criterion.

So that, we choose the second model: included MYCT, MMIN, MMAX, CACH, and CHMAX.

4.2.4 Linear equation

After applying Stepwise regression and assumption check, we have our final multiple linear regression model as concluded. The regression coefficients for each independent variable in this model are as follows:

Result

```
PRP = -56.08 + 0.04911 * MYCT + 0.01518 * MMIN + 0.005562 * MMAX
+ 0.6298 * CACH + 1.46 * CHMAX
```

Figure 21: The regression coefficients

To conclude, the multiple linear regression models provided suggest that variables including MYCT, MMIN, MMAX, CACH, and CHMAX may have a significant effect on the PRP, as indicated by the high R-squared value of 0.8615.

4.2.5 Model Prediction

Now, let's test the equation in **Figure 21** to predict the PRP values in the testing data set that we have splitted above.

Result

index	testing_PRP	predicted_PRP
17	19	-1.8586701
24	76	124.3006081
42	40	78.0650077
47	18	0.8249366
50	62	22.7303861
51	24	30.0861126
70	32	34.4652498
72	64	112.7550649
75	44	25.8182263
78	53	70.0434067

Figure 22: Comparison of testing and predicting value

Some of the observations:

1. For the first observation (index 17), the actual PRP is 19, but the model underestimates it, predicting a value of approximately -1.86.
2. In the second observation (index 24), the actual PRP is 76, whereas the model overestimates it significantly with a prediction of approximately 124.3.

3. For the observation at index 47, the model performs well. The actual PRP is 18, and the model predicts a value very close to it, approximately 0.825.

4. In the observation at index 51, the actual PRP is 24, and the model overestimates it with a prediction of approximately 30.086.

5. In the case of observation index 72, the actual PRP is 64, while the model predicts a value of approximately 112.75, overestimating the PRP.

From these observations, it seems that there is a discrepancy between the model's predictions and the actual PRP values, with the model often overestimating the PRP. Further model tuning and validation could be necessary to improve these predictions.

4.3 Summary

We use multivariate linear regression model evaluated six predictors for PRP: MYCT, MMIN, MMAX, CACH, CHMIN, and CHMAX. However, only five, excluding CHMIN because the R-squared of model which excluding CHMIN is the largest.
These variables accounted for approximately 86.15% of the variation in PRP, as suggested by the adjusted R-squared value. The stepwise regression further confirmed this, resulting in an improved model. The model's prediction quality varied, indicating potential need for further tuning. The final model equation was identified, highlighting the coefficients of the significant predictors.

5 Discussion and Extension

5.1 Linear Regression

5.1.1 Advantages

The biggest advantage of linear regression models is linearity: It makes the estimation procedure simple and, most importantly, these linear equations have an easy to understand interpretation on a modular level. The mathematical equation of Linear Regression is also fairly easy to understand and interpret.

5.1.2 Disadvantages

5.1.2.a Prone to be underfitting

Underfitting is a sitiuation that arises when a machine learning model fails to capture the data properly.This typically occurs when the hypothesis function cannot fit the data well.
In most real life scenarios the relationship between the variables of the dataset isn't linear and hence a straight line doesn't fit the data properly. In such situations a more complex function can capture the data more effectively.Because of this most linear regression models have low accuracy.

5.1.2.b Sensitive to outliers

Outliers of a data set are anomalies or extreme values that deviate from the other data points of the distribution.Data outliers can damage the performance of a machine learning model drastically and can often lead to models with low accuracy.

5.2 Extension: Polynomial Regression

In general, we can model the expected value of y as an nth degree polynomial, yielding the general polynomial regression model

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \cdots + \beta_n x^n + \varepsilon \ .$$

5.2.1 Quadratic Polynomial

Result

```
Residuals:
     Min       1Q   Median       3Q      Max
-154.031  -13.896   -1.474   13.541  192.797

Coefficients:
                Estimate Std. Error t value Pr(>t)
(Intercept)       105.62       2.85  37.065  < 2e-16 ***
poly(MYCT, 2)1    -13.03      50.27  -0.259 0.795706
poly(MYCT, 2)2    -20.96      46.42  -0.451 0.652146
poly(MMIN, 2)1    690.91      82.62   8.363 1.13e-14 ***
poly(MMIN, 2)2    144.77      56.86   2.546 0.011658 *
poly(MMAX, 2)1    908.53      85.21  10.662  < 2e-16 ***
poly(MMAX, 2)2    459.58      64.34   7.143 1.75e-11 ***
poly(CACH, 2)1    516.39      60.22   8.575 2.98e-15 ***
poly(CACH, 2)2   -190.04      49.72  -3.822 0.000178 ***
poly(CHMIN, 2)1   166.60      65.80   2.532 0.012122 *
poly(CHMIN, 2)2  -105.69      52.87  -1.999 0.046984 *
poly(CHMAX, 2)1   276.93      64.01   4.326 2.41e-05 ***
poly(CHMAX, 2)2   300.16      57.79   5.194 5.14e-07 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 41.2 on 196 degrees of freedom
Multiple R-squared:  0.9382,    Adjusted R-squared:  0.9344
F-statistic: 247.8 on 12 and 196 DF,  p-value: < 2.2e-16
```

Figure 23: Quadratic Polynomial result

This output displays the results of a multivariate linear regression where the 'PRP' target variable is modeled as a second-degree polynomial function of the variables 'MYCT', 'MMIN', 'MMAX', 'CACH', 'CHMIN', and 'CHMAX'.

The model's 'Multiple R-squared' is 0.9382, indicating that about 93.82% of the 'PRP' variation can be explained by the selected variables. The F-statistic of 247.8 with a near-zero p-value suggests that at least one predictor is significantly related to 'PRP'.

5.2.2 Cubic Polynomial

Result

```
Call:
lm(formula = PRP ~ poly(MYCT, 3) + poly(MMIN, 3) + poly(MMAX,
    3) + poly(CACH, 3) + poly(CHMIN, 3) + poly(CHMAX, 3), data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-175.908  -15.461   -1.398   12.269  165.832

Coefficients:
                 Estimate Std. Error t value Pr(>t)
(Intercept)      105.6220     2.4325  43.421  < 2e-16 ***
poly(MYCT, 3)1     0.5337    49.0084   0.011  0.99132
poly(MYCT, 3)2    -1.7543    42.4878  -0.041  0.96711
poly(MYCT, 3)3     1.5421    41.5397   0.037  0.97043
poly(MMIN, 3)1   768.2533    73.4441  10.460  < 2e-16 ***
poly(MMIN, 3)2   286.6572    59.3583   4.829 2.81e-06 ***
poly(MMIN, 3)3   -71.3107    42.4813  -1.679  0.09487 .
poly(MMAX, 3)1   724.1310    81.5970   8.874 5.20e-16 ***
poly(MMAX, 3)2   205.0179    62.2989   3.291  0.00119 **
poly(MMAX, 3)3   -77.4150    42.9495  -1.802  0.07306 .
poly(CACH, 3)1   495.6065    53.0409   9.344  < 2e-16 ***
poly(CACH, 3)2  -139.3870    44.8703  -3.106  0.00218 **
poly(CACH, 3)3  -120.2922    41.0113  -2.933  0.00377 **
poly(CHMIN, 3)1  192.1522    59.7388   3.217  0.00153 **
poly(CHMIN, 3)2   55.1256    52.8967   1.042  0.29867
poly(CHMIN, 3)3  -23.9528    46.9365  -0.510  0.61042
poly(CHMAX, 3)1  444.2613    60.9886   7.284 8.38e-12 ***
poly(CHMAX, 3)2  381.4948    52.4464   7.274 8.90e-12 ***
poly(CHMAX, 3)3  415.4717    55.9665   7.424 3.72e-12 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 35.17 on 190 degrees of freedom
Multiple R-squared:  0.9563,    Adjusted R-squared:  0.9522
F-statistic: 231.1 on 18 and 190 DF,  p-value: < 2.2e-16
```

Figure 24: Cubic Polynomial result

The output shows a multivariate linear regression result where the 'PRP' target variable is modeled as a third-degree polynomial function of the variables 'MYCT', 'MMIN', 'MMAX', 'CACH', 'CHMIN', and 'CHMAX'.

The model's 'Multiple R-squared' value of 0.9563 indicates that approximately 95.63% of the 'PRP' variation can be accounted for by the chosen variables.

5.2.3 Quartic Polynomial

Result

```
lm(formula = PRP ~ poly(MYCT, 4) + poly(MMIN, 4) + poly(MMAX,
    4) + poly(CACH, 4) + poly(CHMIN, 4) + poly(CHMAX, 4), data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-129.872  -13.065   -1.553   12.648  155.248

Coefficients:
                 Estimate Std. Error t value Pr(>t)
(Intercept)       105.622      2.389  44.205  < 2e-16 ***
poly(MYCT, 4)1    -51.555     53.387  -0.966  0.33547
poly(MYCT, 4)2     17.304     43.867   0.394  0.69370
poly(MYCT, 4)3     -8.480     42.054  -0.202  0.84041
poly(MYCT, 4)4     36.736     39.209   0.937  0.35003
poly(MMIN, 4)1    747.791     74.414  10.049  < 2e-16 ***
poly(MMIN, 4)2    406.178     75.246   5.398 2.06e-07 ***
poly(MMIN, 4)3     -8.792     52.552  -0.167  0.86731
poly(MMIN, 4)4     -7.422     41.449  -0.179  0.85808
poly(MMAX, 4)1    741.115     82.662   8.966 3.47e-16 ***
poly(MMAX, 4)2    163.162     63.900   2.553  0.01148 *
poly(MMAX, 4)3    -87.932     44.962  -1.956  0.05202 .
poly(MMAX, 4)4    -39.857     37.472  -1.064  0.28888
poly(CACH, 4)1    489.249     54.435   8.988 3.01e-16 ***
poly(CACH, 4)2   -120.157     45.015  -2.669  0.00828 **
poly(CACH, 4)3   -130.349     41.833  -3.116  0.00213 **
poly(CACH, 4)4     76.032     38.691   1.965  0.05090 .
poly(CHMIN, 4)1   170.930     62.264   2.745  0.00665 **
poly(CHMIN, 4)2     1.108     56.180   0.020  0.98429
poly(CHMIN, 4)3    57.059     54.917   1.039  0.30016
poly(CHMIN, 4)4   107.920     61.139   1.765  0.07920 .
poly(CHMAX, 4)1   424.091     62.071   6.832 1.18e-10 ***
poly(CHMAX, 4)2   396.313     55.403   7.153 1.94e-11 ***
poly(CHMAX, 4)3   402.447     55.528   7.248 1.13e-11 ***
poly(CHMAX, 4)4    80.456     46.610   1.726  0.08600 .
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 34.54 on 184 degrees of freedom
Multiple R-squared:  0.9592,    Adjusted R-squared:  0.9539
F-statistic: 180.2 on 24 and 184 DF,  p-value: < 2.2e-16
```

Figure 25: Quartic Polynomial result

This output presents the results of a multivariate linear regression where the target variable 'PRP' is modeled as a fourth-degree polynomial function of the variables 'MYCT', 'MMIN', 'MMAX', 'CACH', 'CHMIN', and 'CHMAX'.

Finally, the model's 'Multiple R-squared' is 0.9592, suggesting that about 95.92% of the variation in 'PRP' can be explained by the selected variables. The 'Adjusted R-squared' is slightly lower, as it adjusts for the number of predictors in the model, penalizing the addition of uninformative predictors.

The F-statistic is 180.2 with a p-value close to zero, indicating that at least one of the predictors is significantly related to the outcome variable at the 5% significance level.

5.2.4 Polynomial regression coefficients

Due to the highest R - Square, we choose the Quartic Polynomial equations

Result

```
PRP = 105.622010 - 51.554743 * poly(MYCT, 4)1 + 17.303600 * poly(MYCT, 4)2
- 8.480363 * poly(MYCT, 4)3 + 36.735613 * poly(MYCT, 4)4 + 747.790838
poly(MMIN, 4)1 + 406.178362 * poly(MMIN, 4)2 - 8.792411 * poly(MMIN, 4)3
- 7.422042 * poly(MMIN, 4)4 + 741.114949 * poly(MMAX, 4)1 + 163.162008 *
poly(MMAX, 4)2 - 87.932231 * poly(MMAX, 4)3 - 39.857313 * poly(MMAX, 4)4
+ 489.249181 * poly(CACH, 4)1 - 120.157204 * poly(CACH, 4)2 - 130.348582
* poly(CACH, 4)3 + 76.032362 * poly(CACH, 4)4 + 170.929558 * poly(CHMIN, 4)1
+ 1.107751 * poly(CHMIN, 4)2 + 57.059371 * poly(CHMIN, 4)3 + 107.920488
* poly(CHMIN, 4)4 + 424.090643 * poly(CHMAX, 4)1 + 396.312629
* poly(CHMAX, 4)2 + 402.447204 * poly(CHMAX, 4)3 + 80.456002 * poly(CHMAX, 4)4
```

Figure 26: The Quartic regression coefficients

Result

index	testing_PRP	predicted_PRP
17	19	31.96338
24	76	44.28087
42	40	58.80600
47	18	11.57624
50	62	35.85962
51	24	32.16911
70	32	41.97279
72	64	53.84509
75	44	44.63116
78	53	55.40016

Figure 27: Prediction on testing set by chosen polynomial equation above

Some of the observations:

1. For the first observation (index 17), the actual PRP is 19, but the model overestimates it, predicting a value of approximately 31.96338.
2. In the second observation (index 24), the actual PRP is 76, whereas the model underestimates it with a prediction of approximately 44.28087.
3. For the observation at index 47, the model performs well. The actual PRP is 18, and the model predicts a value very close to it, approximately 11.57624.
4. In the observation at index 51, the actual PRP is 24, and the model overestimates it with a prediction of approximately 32.16911.
5. In the case of observation index 72, the actual PRP is 64, while the model predicts a value of approximately 53.84, overestimating the PRP.

From these observations, even though, there is still some problem in prediction values, it seems that we have obtained a better result .

5.2.5 Model comparison

To determine between linear regression and quartic polynomial regression which one is more efficient, we will consider the rate of accuracy of two models(calculated by subtracting SSE from SST, dividing the result by SST, and then multiplying by 100 to express the accuracy as a percentage)

Result

```
The accuracy of the linear regression model on test set:  91.16 %
The accuracy of the quartic polynomial regression model on test set:
95.44 %
```

The quartic polynomial regression model is obviously more efficient than linear regression model because of greater accuracy rate.

6 Code and data source

6.1 Code source

Link: Code

6.2 Data source

Link: Data

References

[1] Douglas C. Montgomery, *Applied Statistics and Probability for Engineers*

[2] FAIZUNNABI, *Comp Hardware Performance*, URL: https://www.kaggle.com/datasets/faizunnabi/comp-hardware-performance?select=machine.data.txt

[3] Antoine Soetewey, *Descriptive statistics in R*, URL: https://statsandr.com/blog/descriptive-statistics-in-r/density-plot

[4] John Verzani, *simpleR - Using R for Introductory Statistics*

[5] Newcastle University. (n.d.). *Coefficient of Determination, R-squared*. Retrieved November 21, 2022, URL: https://www.ncl.ac.uk/webtemplate/ask-assets/external/maths-resources/statistics/regression-and-correlation/coefficient-of-determination-r-squared.html