Problem Solving Task 2

MXN500

Due date	31/05/2020 11:59pm
Last name	Nayak
First name	Drashti
Student number	N10599568

About this Problem Solving Task

On completion of this unit you should be able to analyse a data set relating to business, IT, or one of the natural sciences, estimate parameters in an empirical relationship and make inferences based on model diagnostics.

Here's a checklist of the steps you need to go through before submitting this online.

- **Individualise your data** by removing the rows with your student number.
- **Answer the questions** in the relevant areas in this document.
  - For all steps that require you to show your working, give enough detail (and explanatory text where needed) that anyone reading can follow the logic of your solution.
  - Correct answers with no working will not attract full marks.
  - Provide any R code used to solve the problem in your R script file
- **Interpret the analysis** in the space provided in the Interpretation section
  - Ensure all plots are included at a resolution that makes them readable and that they have either a caption or a title
- **Provide your completed document** by uploading it to Blackboard as either a Microsoft Word file of format .doc or .docx or a PDF file of format .pdf. Pages files will be awarded a mark of 0
- **Provide your R script** file
  - PSTs with no script files will be considered incomplete and awarded a mark of 0
  - The script file should
    - have file extension `.R` or `.rmd`
    - contain all code required to read in the downloaded data, and
    - perform all the steps of the analysis that require the use of R

Submit your PST and R script file on Blackboard by the due date. Ensure you have attached both files.

This is individual assessment. You should not work with others to complete the task, share your work, or ask other students for their work.

Your solution must be your own. You are not permitted to copy, summarise, or paraphrase the work of others in your solution.

Please ensure you are familiar with QUT's policy on Academic Integrity

Table of Contents

Introduction

For this Problem Solving Task we will be investigating the advance of electronics. In computing, Moore's law, which originated in around 1970, is a rule of thumb stating that overall processor speeds will double every two years. This can be re-stated, more specifically, as a predicted doubling of the clock speed on an affordable CPU every two years. Your `PST2data.csv` file contains information gathered from 1970 onwards on the number of transistors, the clock speeds, power density and number of cores of selected "affordable" CPUs. Your task will be to use your knowledge of linear regression to find an equation relating clock speed to the time since 1970.

Setup

Ensure you filter the data to remove any rows with your student number before completing the analysis. Once you have subsetted by student number, remove the variable `Student` from the data frame.

Section A - Data and Visualisation (10%)

A1 Data Structure (5%)

Before we can start fitting models, we need to get a clear picture of our data.

Exercise: Produce a captioned, well-formatted table containing a data dictionary of the variables in PST2data.csv. Include the variable name, type (nominal, ordinal, interval, ratio, count, etc), units, and typical range of the variable. *Note: You may wish to do some research to find information like units.*

Answer:

	Variable	type	units	Range
1	Year	nominal	time	1971-2015
2	Transistors	interval	nanometer(nm)	2.30e+03-5.56e+09
3	Clock..MHz.	ratio	MHz(1 million cycles per second)	0.74-3730.00
4	Power.Density	ratio	Volume(W/mcube)	1.18125-106.17284
5	Cores	Count	no.of cores(amount)	1-18

Table 1: containing a data dictionary of the variables

Exercise: Transform your data to include a column `TimeSince` which is the time in years since 1970. Include the code you used here.

Answer:

year = 1970

mydf$timeSince = (mydf$Year - year)

Exercise: Generate a captioned, well-formatted table that shows the minimum, median, and maximum of the number of transistors, clock, power density, and cores.

Answer:

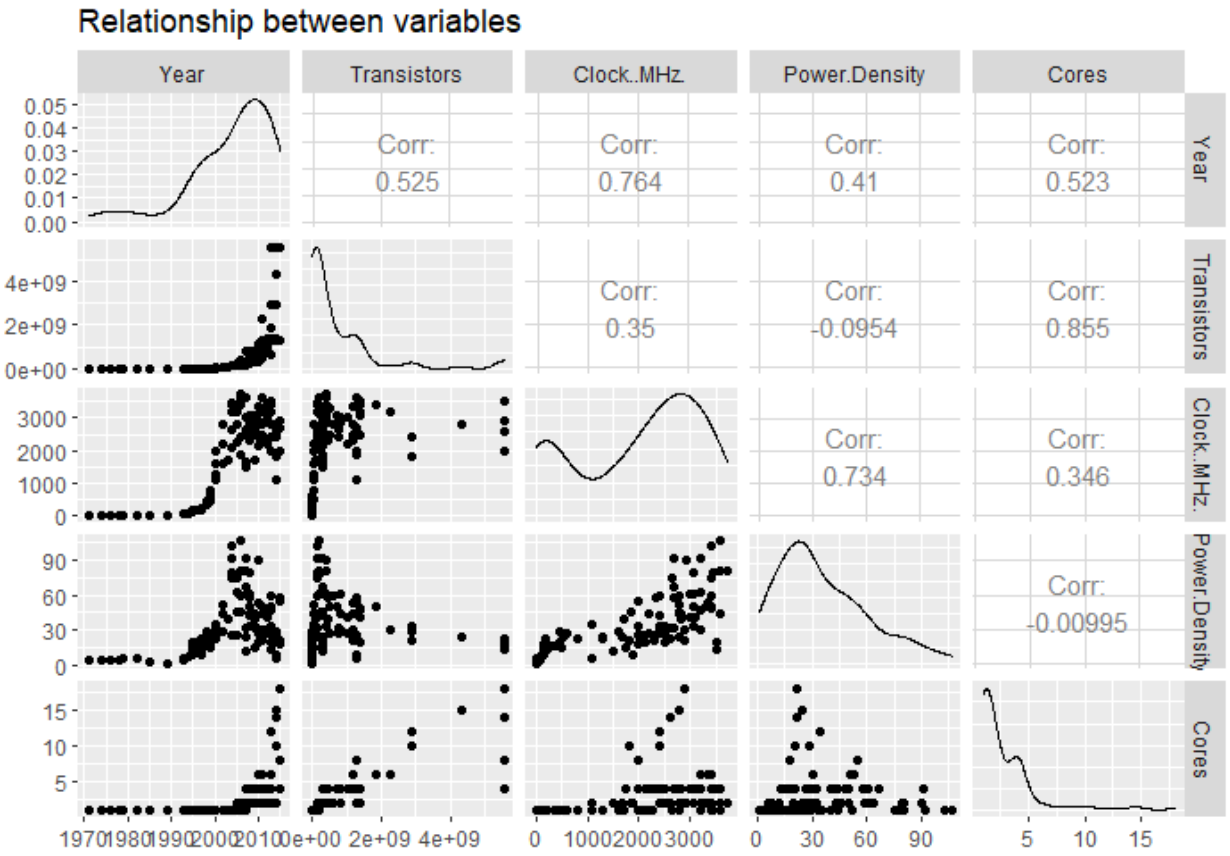
	Stats	transistor	clock	Power.Density	Cores
1	Minimum	2.300e+03	0.74	1.181	1.000
2	Median	2.605e+08	2400.00	28.664	2.000
3	Maximum	5.560e+09	3730.00	106.173	18.000

Table 2: Minimum, median and maximum of transistors, clock, power density and cores.

A2 Graphical Summaries (5%)

Exercise: Using the `ggpairs` function in the `GGally` package, create a pairwise plot showing the relationship between each variable in the dataset.

Answer:



Transistors and cores are highly correlated.

Fig.1 the relationship between each variable using ggpair function

Exercise: Create a graphically excellent plot showing the variablility in the clock speed over time since 1970. Include a straight line of best fit and ensure your plot has appropriate labels. **Hint**: You should transform one or both of your variables using a natural logarithm such that your data is close to the straight line of best fit.

Answer: With log on clock speed the plot of data is closes to best fit straight line.

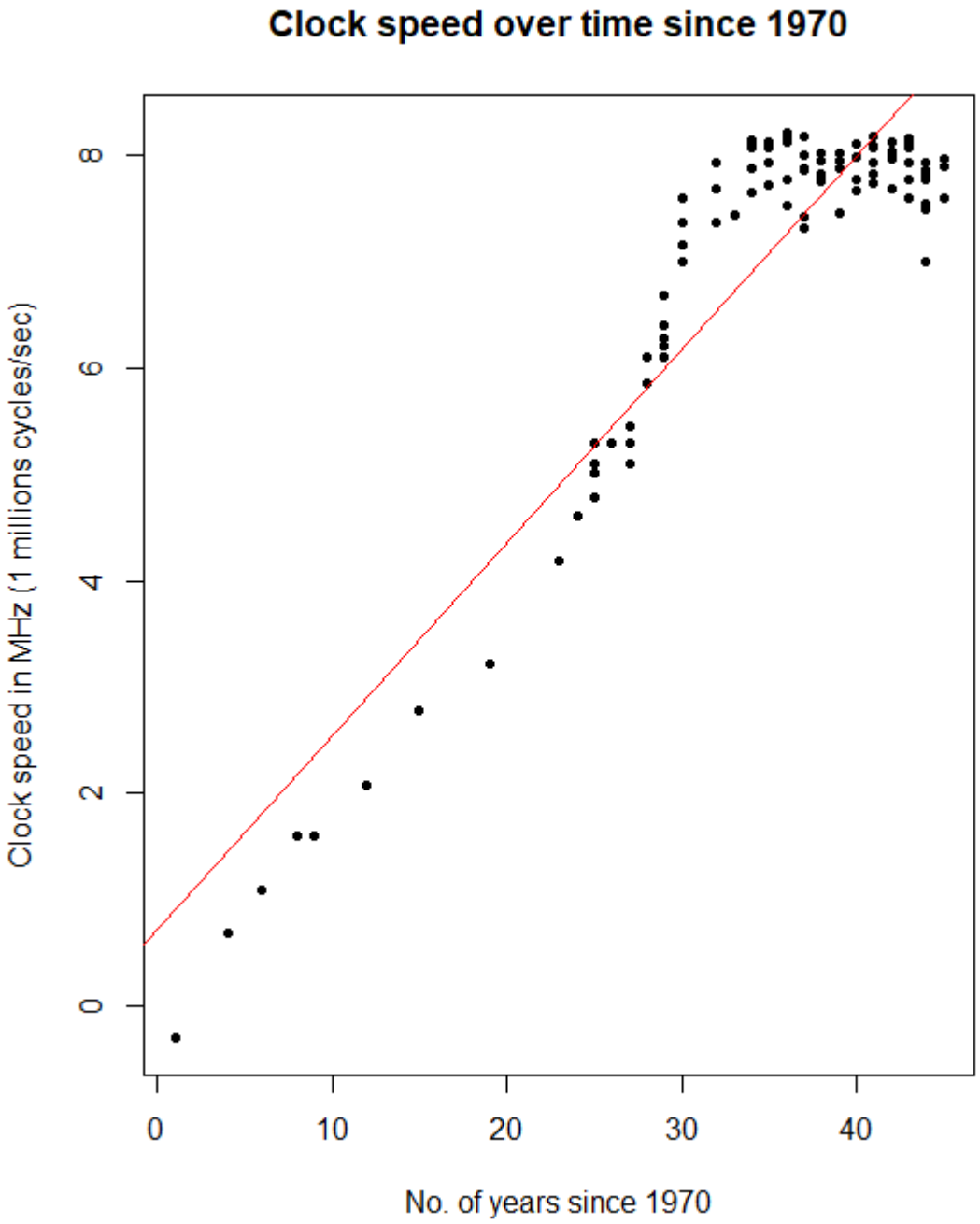


Fig. 2 the variablility in the clock speed over time since 1970.

Section B - Linear Regression (30%)

B1 Setup (10%)

Given that overall processor speed directly relates to the clock speed (MHz), we must decide which of the above relationships to investigate further. Sometimes it may be necessary to take transformations of one or both variables in order to linearise the data.

Exercise: From the plot you generated relating the clock speed to the time since 1970, which combination of transformations linearises the data most effectively?

Answer: with log on only clockspeed and not on timesince the data smoothly linearises.

We aim to fit a linear regression model to estimate the relationship between clock speed and time.

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2)$$

{#eq:eqn1}

where the $\varepsilon_i \sim N(0, \sigma^2)$ are the errors/residuals, distributed normally with a mean of 0 and standard deviation σ.

Exercise: Complete the following sentence so that it refers to the terms in the regression model above. You must include the phrase "the natural log of" to indicate which variable(s) you transformed.

Answer: A linear model was specified to model how the clock speed to the natural log of y, is related to the time since 1970,x. The parameter β_1 describes the rate of change of the clock speed, and the parameter β_0 represents the clock speed in CPUs available in 1970.

B2 Linear Model (10%)

Exercise: Using R, fit a linear model to your chosen variables. Produce a captioned, well-formatted table below that includes a descriptive parameter name, estimate and 95% confidence interval.

Answer:

```
Call:
lm(formula = log(Clock..MHz.) ~ timeSince, data = lr_df)

Residuals:
     Min       1Q   Median       3Q      Max
-1.71683 -0.47629 -0.05038  0.49946  1.42784

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.715545   0.261994   2.731  0.00742 **
timeSince   0.181917   0.007417  24.528  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7163 on 104 degrees of freedom
Multiple R-squared:  0.8526,    Adjusted R-squared:  0.8512
F-statistic: 601.6 on 1 and 104 DF,  p-value: < 2.2e-16
```

Table 3: a descriptive parameter name, estimate and 95% confidence interval of linear model.

Exercise: Based on your parameter estimates, write down your linear model as per equation (1).

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,$$

Answer: $\log(\hat{clockspeed}) = 0.716 + (0.182)\ timeSince + \varepsilon_i$

```
> get_regression_table(linmod)
# A tibble: 2 x 7
  term      estimate std_error statistic p_value lower_ci upper_ci
  <chr>        <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
1 intercept    0.716     0.262      2.73   0.007    0.196     1.24
2 timeSince    0.182     0.007     24.5    0        0.167     0.197
>
```

Fig 3: estimates output for linear model.

Exercise: How much of the variability in the observed data does the model explain?

Answer: the variability in the observed data is 85.26%

B3 Analysis of Residuals (10%)

Whilst summary statistics can be good indicators of model fit, other methods must be used to ensure the underlying assumptions of linear regression are met.

Exercise: Create a plot that shows how the residuals vary with the values fitted through your linear regression model.

Answer:

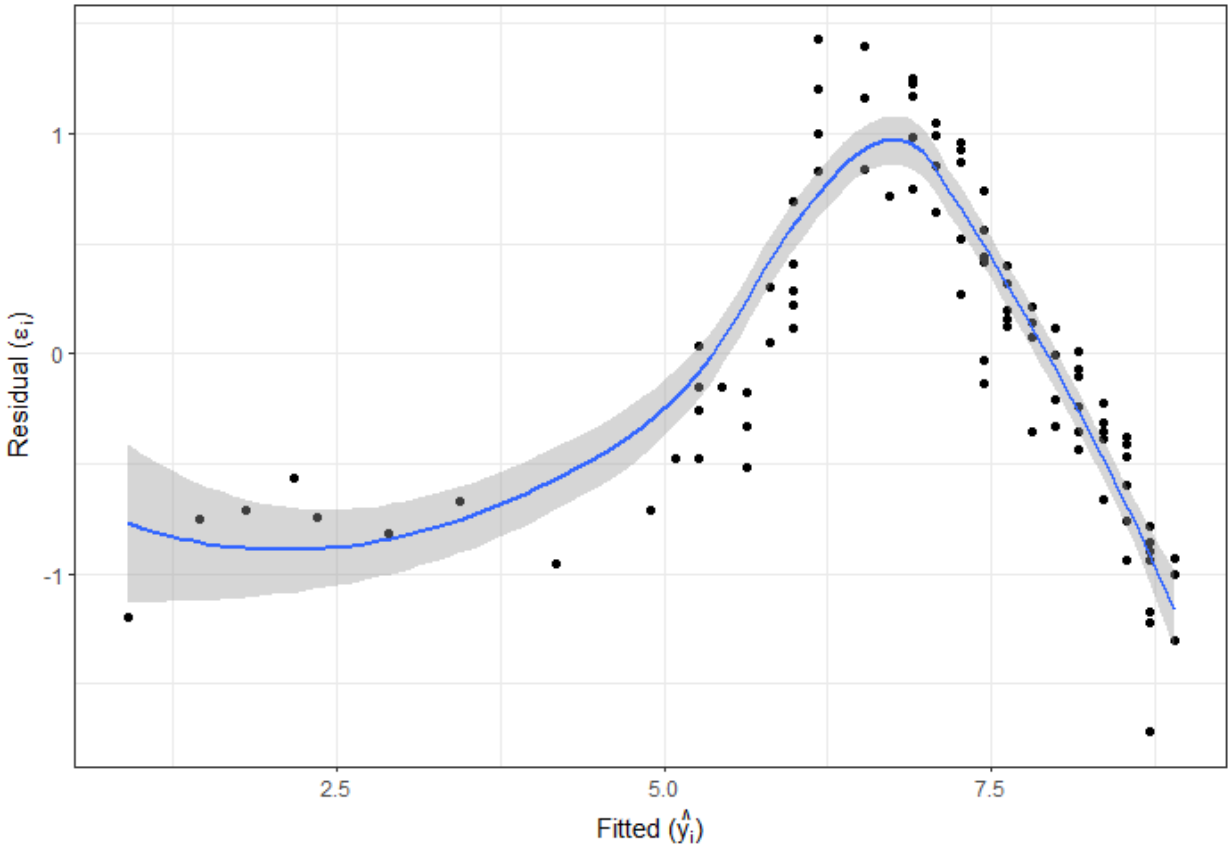


Fig 4: shows how the residuals vary with the values fitted through your linear regression model

Exercise: Create a plot that shows how the residuals vary with the power density.

Answer:

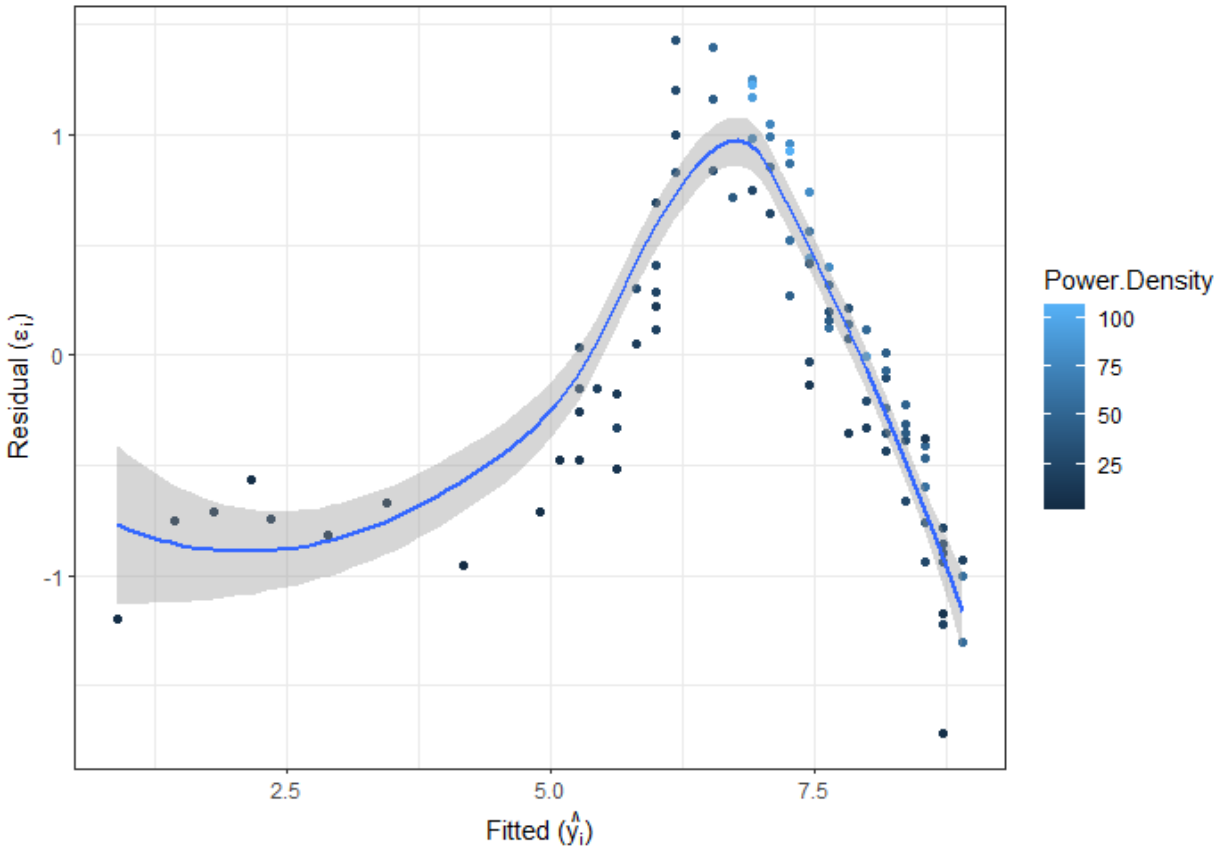


Fig 5: show the residuals vary with the power density

Exercise: Create a QQ plot that compares the standardised residuals to a standard normal distribution.

Answer:

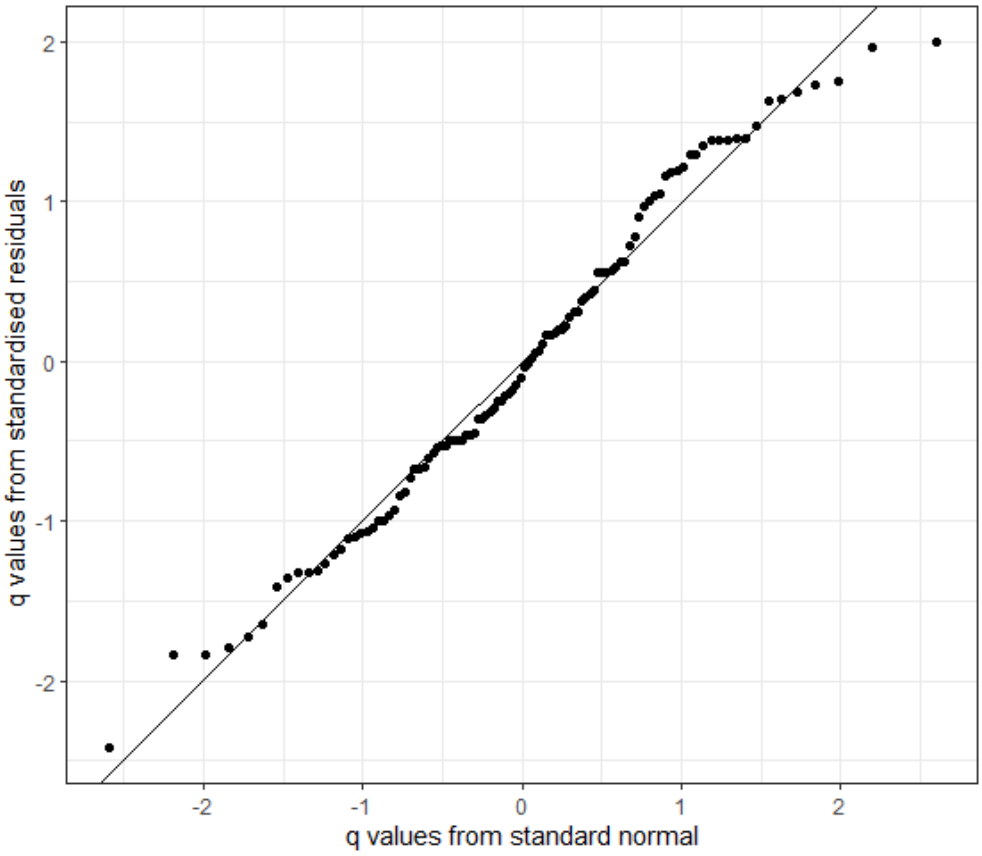


Fig 6: comparison of the standardized residuals to a standard normal distribution.

Section C - Advanced Regression (40%)

In this section, you will modify your original model to include an explanatory term (Power Density).

C1 Multiple Explanatory Variables (10%)

We will use the form

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 p_i + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2)$$

{#eq:eqn2}

where the $\varepsilon_i \sim N(0, \sigma^2)$ are the errors/residuals, distributed normally with a mean of 0 and standard deviation σ.

Exercise: Using R, fit a linear model to your chosen variables. Produce a captioned, well-formatted table below that includes a descriptive parameter name, estimate and 95% confidence interval.

Answer:

```
Call:
lm(formula = log(Clock..MHz.) ~ timeSince + Power.Density, data = mydf)

Residuals:
     Min       1Q   Median       3Q      Max
-1.45519 -0.34932  0.01853  0.30312  1.39034

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.734171   0.201191   3.649 0.000415 ***
timeSince     0.159984   0.006244  25.621  < 2e-16 ***
Power.Density 0.020457   0.002388   8.566 1.12e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.55 on 103 degrees of freedom
Multiple R-squared:  0.9139,    Adjusted R-squared:  0.9123
F-statistic: 546.9 on 2 and 103 DF,  p-value: < 2.2e-16
```

Table 4: a descriptive parameter name, estimate and 95% confidence interval of linear model with explanatory variable.

Exercise: Based on your parameter estimates, write down your linear model as per equation (2).

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 p_i + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2)$$

Answer:
$\log(\text{clockspeed}) = 0.734 + 0.16(\text{timesince}) + 0.02(\text{Power. Density}) + \varepsilon_i$

```
> get_regression_table(fit)
# A tibble: 3 x 7
  term          estimate std_error statistic p_value lower_ci upper_ci
  <chr>            <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
1 intercept        0.734     0.201      3.65       0    0.335    1.13
2 timeSince        0.16      0.006     25.6        0    0.148    0.172
3 Power.Density    0.02      0.002      8.57       0    0.016    0.025
```

Fig 7: estimate output for linear model with explanatory variable.

Exercise: How much of the variability in the observed data does the model explain?

Answer: the variability in the observed data is 91.39%.

C2 Residual Analysis (10%)

Exercise: Create a plot that shows how the residuals vary with the values fitted through your multivariate model.

Answer:

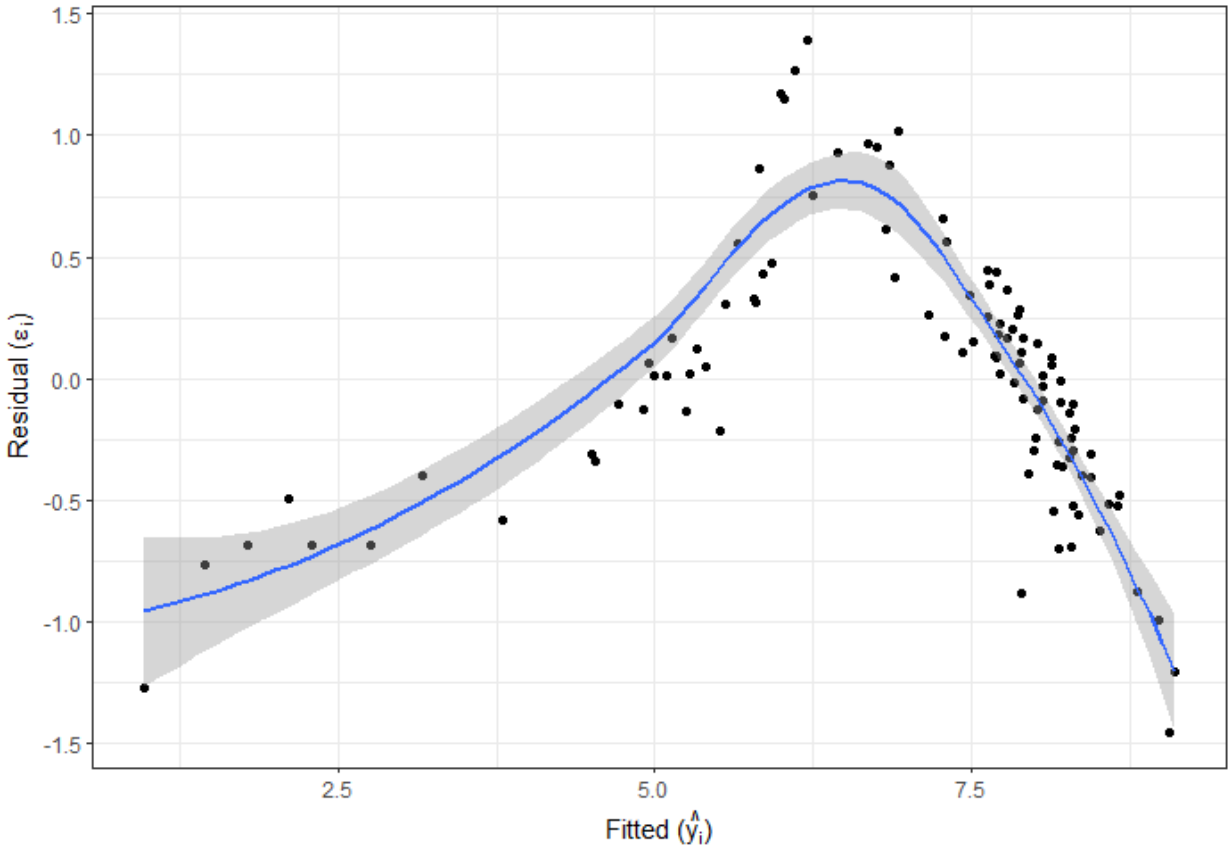


Fig 8: show the residuals vary with the values fitted through your multivariate model.

Exercise: Create a QQ plot that compares the standardised residuals to a standard normal distribution.

Answer:

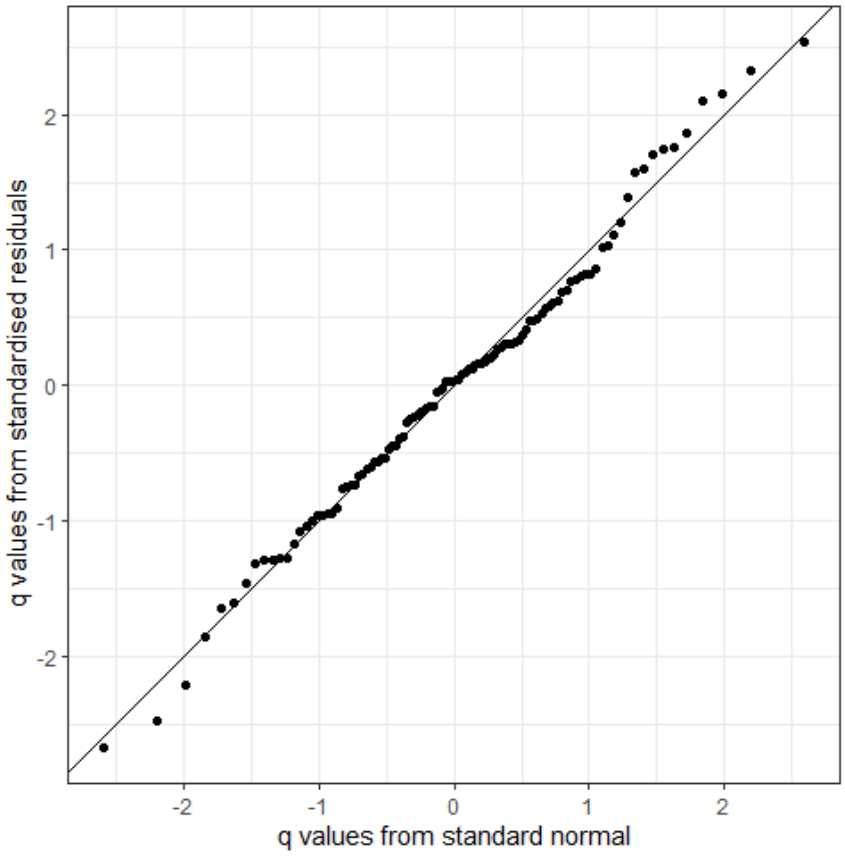


Fig 9: QQ plot that compares the standardised residuals to a standard normal distribution.

C3 Model Choice (10%)

In this section we will perform an F-test for model choice.

Exercise: Identify the full model.

Answer: The full model is taken the model from C1 section where log(clockspeed) is function of timeSince data of years(from 1970) with explanatory variable Power density. The equation of model is as per follows: $\log(\text{clockspeed}) = 0.734 + 0.16(\text{timesince}) + 0.02(\text{Power. Density}) + \varepsilon_i$

Exercise: Identify the reduced model.

Answer: Reduced model basically only contains variables from the full model and those are log(clockspeed) as function of timeSince data of years(from 1970). The equation is as per follows: $\log(\hat{clockspeed}) = 0.716 + (0.182)\ timeSince + \varepsilon_i$

Exercise: State your null and alternative hypothesis.

Answer:

Ho: Amount of variation in reduced model is equal to full model.

H1: Amount of variation in reduced model is less than full model.

Exercise: State the p-value and degrees of freedom for your test.

Answer: P-value of 1.117e-13 and degrees of freedom is 1 for the test. The p value is less than 0.05 hence, we reject the null hypothesis which proves that full model has higher amount of variability.

Exercise: Interpret the results of your test. Which model is best?

Answer: Based on the variability of both the models, model A with explanatory variable is best with 91.39% variability and model B has 85.26% varability. Variability as far away from 0 is better the model. However, I can say that this can be second best model as model with all the variables has higher variability of 92.95%. But as you can see the difference between model with one explanatory variable and model with 5 variables is very small. Hence, sometimes it might result into over fitting which is not good. Hence, full model with explanatory variable power. With Density and time is best among both.

C4 Best Model (10%)

Develop a model to predict clock speed based on any of the variables in the dataset. Be sure to define your model, justify your model choice, and indicate how well it measures the variability in the outcome variable.

For the best model, developed a model to predict natural log of (clock speed) with all variables, time, power density, and cores except transistors because the value of variability doesn't affect much from the dataset. However model with all the variables has higher variability of 92.95%. This includes cores and it's valid because processing speed depends on cores as well, that's multiple CPU's run parallel or individually. And proved with r-squared result. In addition if we remove log of clock speed the data with highest variability and good fit is same as variables considered in above that are power density, time and cores is 81.7% and difference is not much when added transistors i.e. 80.2%. So, it can be concluded that other than time, power density is highly related when it comes to satisfy Moore's law.

In addition, prediction model with power density is also shown as it's highly correlated with clock speed as well. For the best model, developed a model to predict natural log of (clock speed) with power density only from the dataset. The two highest variable are time and density that has shown increasing variability later the model variability is increasing but less interval.

```
> #How much variability in the observed data does both models explain?
> print(paste("Variablitity in full model is", (round(glance(full_model)$r.squared*100,2))))
[1] "Variablitity in full model is 92.95"
> print(paste("Variablitity in reduced model is", (round(glance(fit)$r.squared*100,2))))
[1] "Variablitity in reduced model is 91.39"
>
> anova(full_model, fit)
Analysis of Variance Table

Model 1: log(Clock..MHz.) ~ timeSince + Power.Density + Cores + Transistors
Model 2: log(Clock..MHz.) ~ timeSince + Power.Density
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    101 25.540
2    103 31.161 -2   -5.6207 11.114 4.341e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

Fig 10: Variablity of both model and other parameters for best model.

Section D - Interpretation (20%)

Write 1-3 sentences on each of the following, referring to the results of your analysis where necessary. If you rely on sources of information other than your analysis, ensure you cite them and that a bibliographic entry is provided below. A great response will likely refer to the data, the model, and the real-world.

Exercise: Given the plots you generated throughout this task, are the assumptions for a linear model satisfied? Why or why not?

Answer: There are many constraints to look after when comes to base on Moore's law as Moore's law does not state increase in clock speed , power density or no. of cores. However they are highly correlated and increases with time. And Moore's law has been accurate till now. Based on the analysis and plots released it can be said that clock speed is mostly proportional to power density over time.

Exercise: Based on the analysis completed throughout this PST, is Moore's law a good prediction for the advance of computational power?

Answer: Yes, Moore's law is still valid. It might be closure to its end as according to data the number of transistors are still twice every two years and with some technologies rate is even faster.

Exercise: What are some other variables that may affect the Clock Speed possible?

Answer: The second best suitable variable that affects clocks speed is no. of cores and there is no doubt in why. Multiple cores means multiple CPU's within a single chip which carried out independently or concurrently that affects the processor speed.

Exercise: Looking at the context, and where computers are going, do you think Moore's law will continue to hold for the next decade? Why or why not?

Answer: While looking at the past data of 45 years, Moore's law has almost near to or is resulted on the best fit line which shown in the analysis and neglecting some negative

outliers detected in the process. It does fill that Moore's law will continue to hold for the next decade as well. However, there is a limitation may arise to it as there is a limitation on inserting no. of transistors on a single chip might at one point stop Moore's laws observable trend but that's not easy as well. In many other ways, the speed of processor can be increased and in this fast paced world an innovation is no far way to boost performance.

Refrences:

Wilson, C. (2016, August 10). To me a literature review is about stating what other researchers think of the topic [Comment on the blog post "Literature Review vs. Essay"]. *QUT Library*.

1) Regression diagnostics: testing the assumptions of linear regression , Robert Nau, Duke, University. http://people.duke.edu/~rnau/testing.htm
2) Lecturer, Belinda (2020). *MXN500 Statistical analysis: [Linear Regression, Prediction and model choice, Regression with multiple explanatory variable*]. QUT Blackboard.
3) 360 DataScience Course program, Online, Intro. To R programming and Statistics. https://365datascience.com/
4) Daniel Fishman (2018, September 19). At least until 2030. Moore's law is still holding. [Comment on the blog post " How long will Moore's Law continue to apply?"]. https://www.quora.com/How-long-will-Moores-Law-continue-to-apply