Introduction

Here, 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. The aim is to find an equation relating clock speed to the time since 1970.

Section A - Data and Visualisation

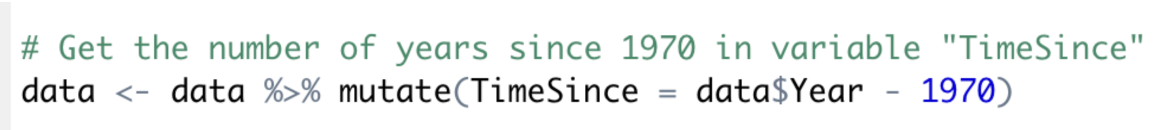
A1 Data Structure

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

1. Data Dictionary

VARIABLE NAME	TYPE	UNITS	TYPICAL RANGE
Year	Nominal	Number of years	1971-2015
Transistors	Count	Number of transistors	2.300e+03 - 5.560e+09
Clock (MHz)	Ratio	MHz	0.74 - 3730
Power Density	Ratio	W/m3	1.181 – 106.173
Cores	Count	Number of cores	1 - 18

2. Transform the data to include a column `TimeSince` which is the time in years since 1970.



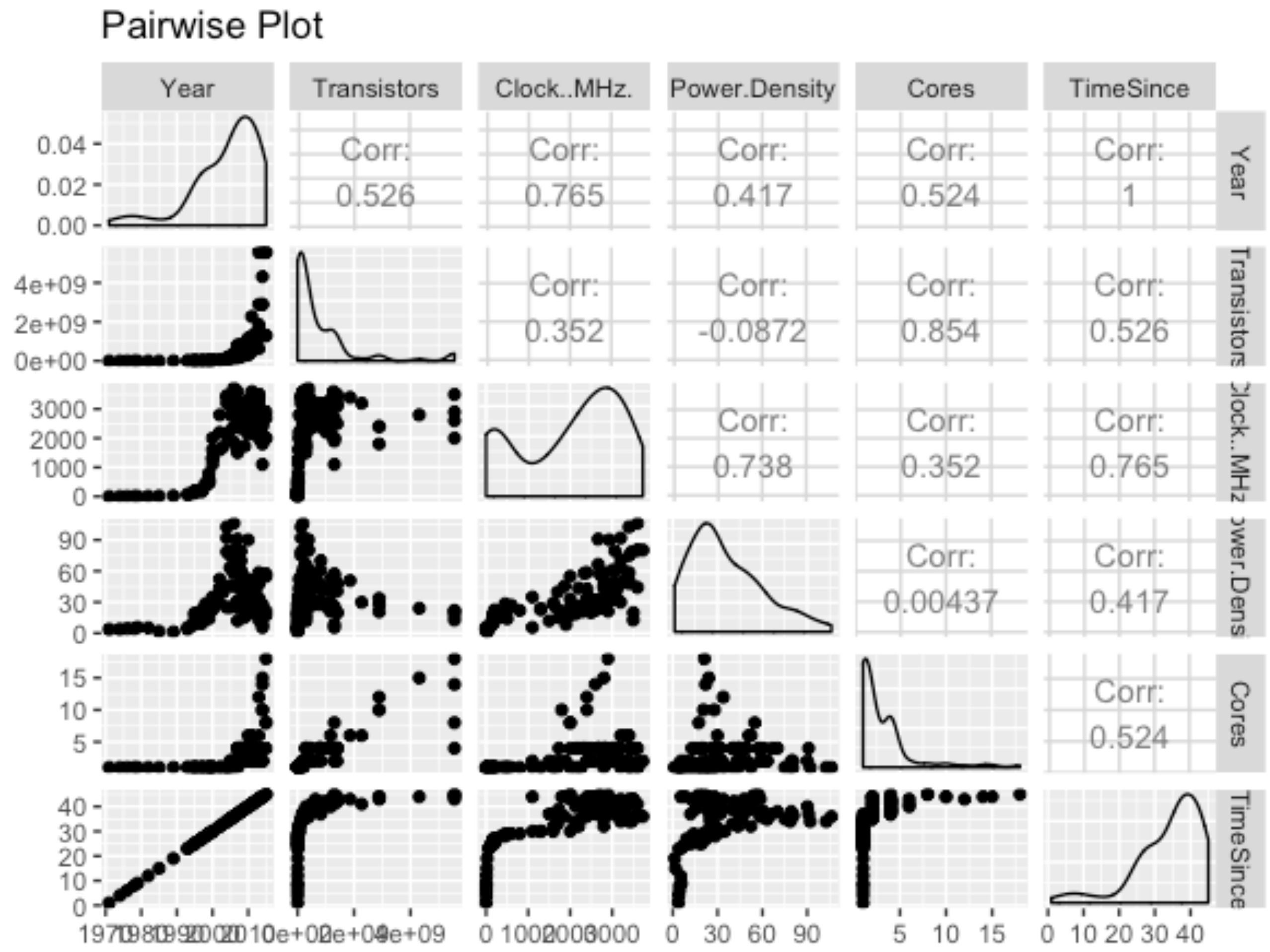
```
# Get the number of years since 1970 in variable "TimeSince"
data <- data %>% mutate(TimeSince = data$Year - 1970)
```

3. Table that shows the minimum, median, and maximum of the number of transistors, clock, power density, and cores.

	Minimum	Median	Maximum
Transistors	2.30e+03	2.91e+08	5.56e+09
Clock (MHz)	0.74	2400.00	3730.00
Power Density	1.181	29.008	106.173
Cores	1.000	2.000	18.000

A2 Graphical Summaries

1. A pairwise plot showing the relationship between each variable in the dataset.



2. A plot showing the variability in the clock speed over time since 1970.

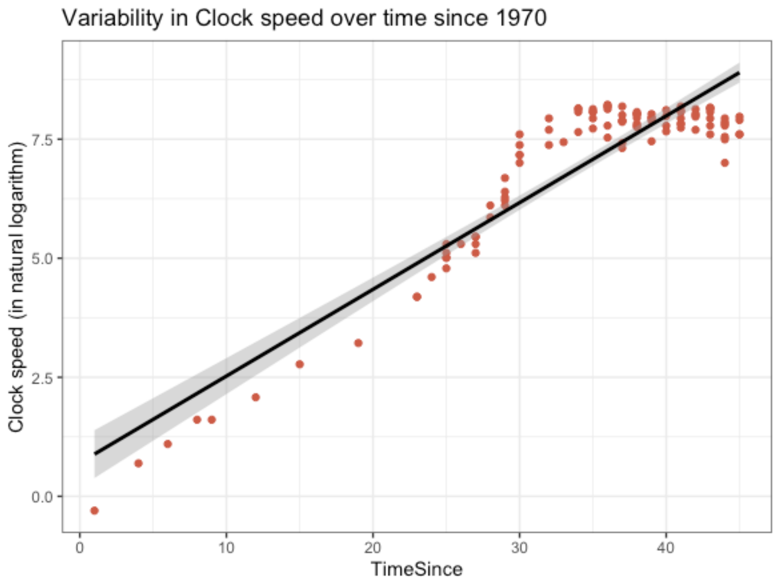


Figure 1Logarithmic scale on y-axis

Section B - Linear Regression

B1 Setup

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.

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

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

$$y_i = 0.704 + 0.182x_i$$

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

As seen from figures 1,2,3 and 4, log(Clock) and TimeSince combination (Logarithmic scale on y-axis) linearises the data most effectively.

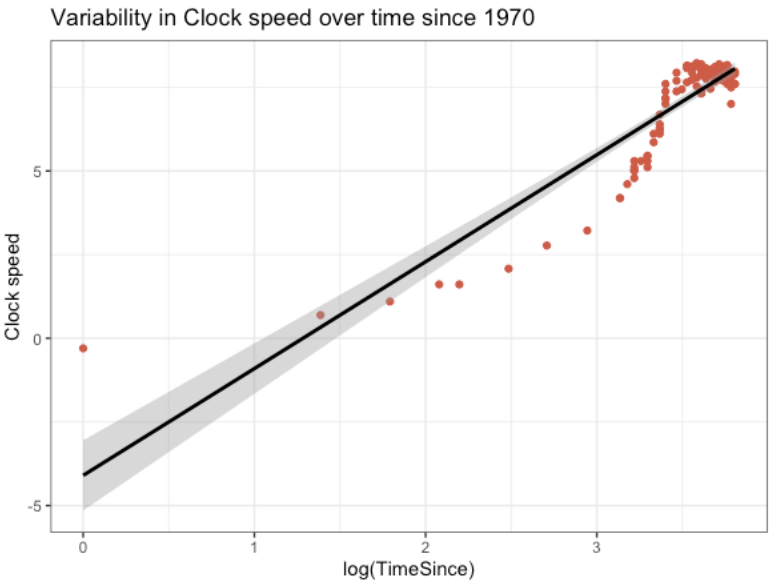


Figure 2Logarithmic scale on both x and y axes

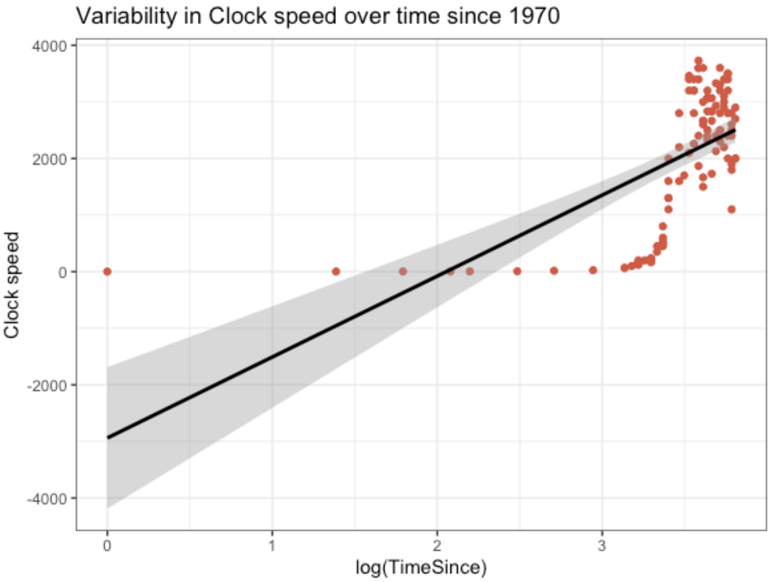


Figure 3 Logarithmic scale on x-axis

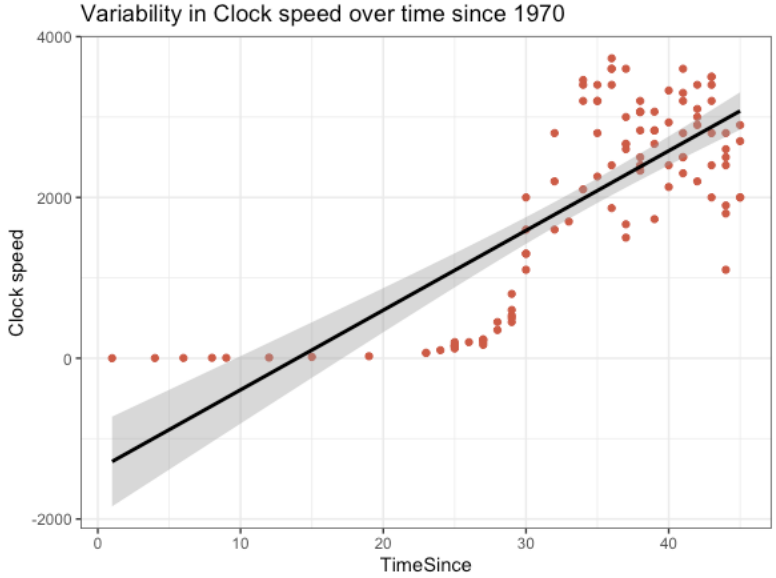


Figure 4 No logarithmic scale

A linear model was specified to model how the clock speed, $y_i(the\ natural \log of\ clock)$, is related to the time since 1970, x_i. 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

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

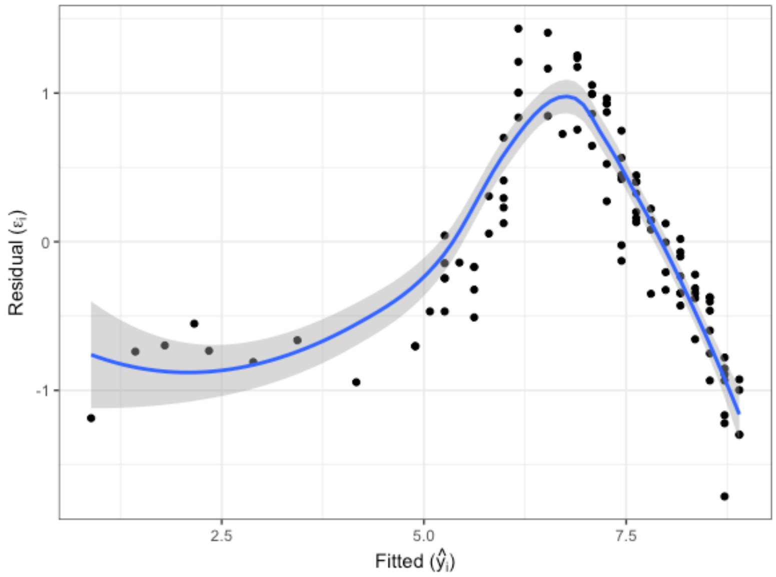
term	estimate	conf.low	conf.high	p.value
(Intercept)	0.704	0.188	1.22	7.98e- 3
TimeSince	0.182	0.168	0.197	2.13e-45

The linear model as per equation (1) is $y_i = 0.704 + 0.182x_i$

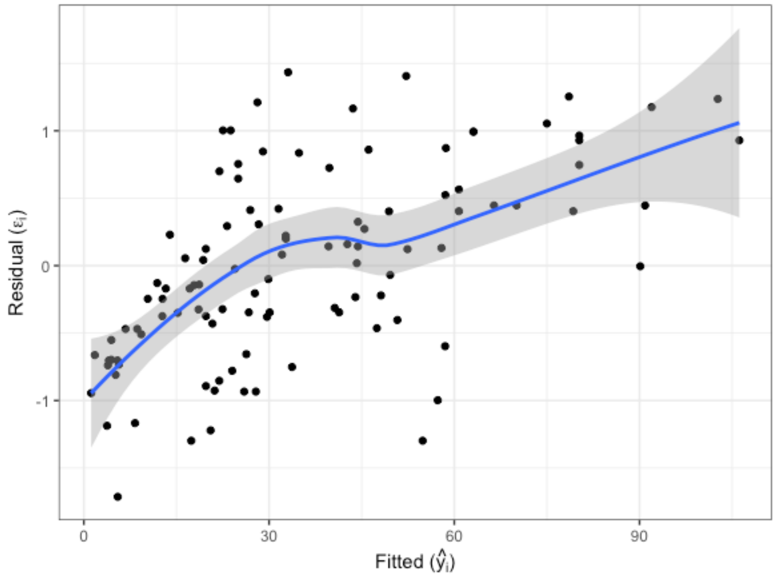
The R-squared value of 0.8549 represents that the variability of the observed data of the departure delay model is 85.49%.

B3 Analysis of Residuals (10%)

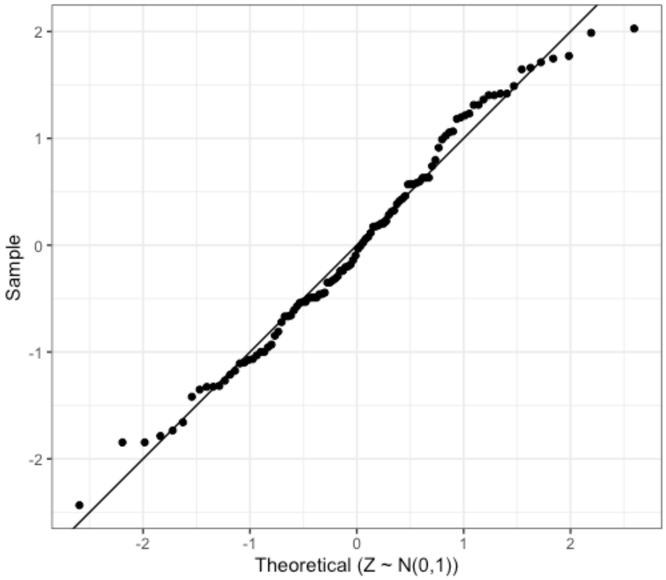
1. A plot that shows how the residuals vary with the values fitted through your linear regression model.



2. A plot that shows how the residuals vary with the power density.



3. A QQ plot that compares the standardised residuals to a standard normal distribution.



Section C - Advanced Regression

In this section, we will modify the 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.

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

term	estimate	conf.low	conf.high	p.value
(Intercept)	0.733	0.334	1.13	4.25e- 4
TimeSince	0.160	0.148	0.173	1.94e-46

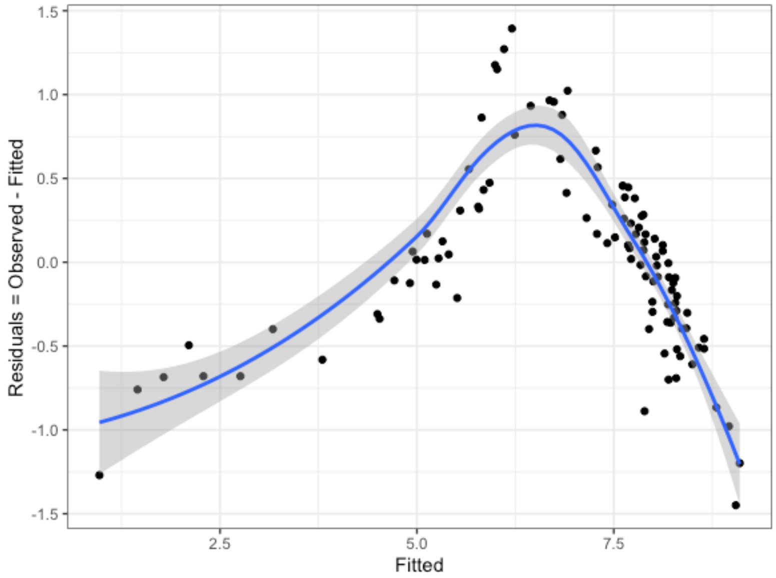
PowerDensity	0.0202	0.0155	0.0250	2.39e-13

Based on the parameter estimates, the linear model as per equation (2) is $y_i = 0.733 + 0.160x_i + 0.0202p_i$

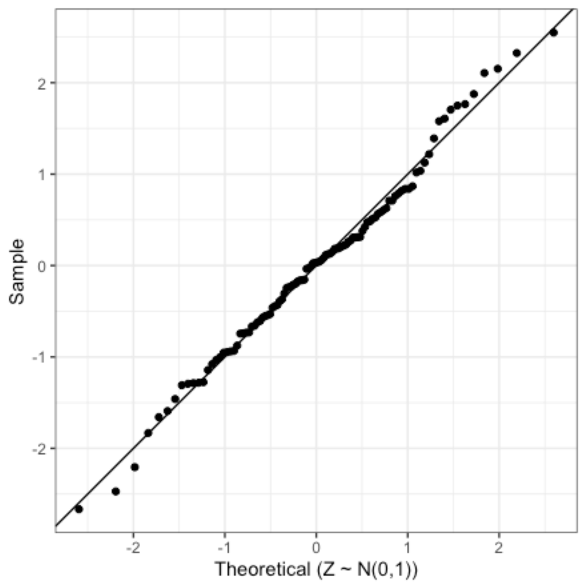
The R-squared value of 0.9140 represents that the variability of the observed data of the departure delay model is 91.40%.

C2 Residual Analysis

1. A plot that shows how the residuals vary with the values fitted through your multivariate model.



2. A 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.

The full model:

Model 2: lClock ~ TimeSince + PowerDensity

The reduced model:

Model 1: lClock ~ TimeSince

H_0: Full model (Model 2) explains no more variability than restricted/reduced model (Model 1)

H_1: Reduced model (Model 1) explains a greater deal of variability.

p-value = 2.386e-13

Degrees of freedom = 1

Since p< 0.05, there is insufficient evidence against the null hypothesis that both the models explain same amount of variation in the data. Therefore, we are in favor of the alternative hypothesis that the second model explains a greater deal of variability.

Therefore, model 2 is better in this analysis.

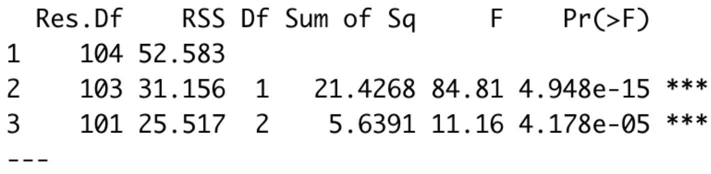
C4 Best Model

Develop a model to predict clock speed based on any of the variables in the dataset.

The models are defined as below:

Model 1: lClock ~ TimeSince
Model 2: lClock ~ TimeSince + PowerDensity
Model 3: lClock ~ TimeSince + PowerDensity + Cores + Transistors

We have fitted the model 3 as a function of transistors, power density, cores and time since with 95% confidence intervals.



```
  Res.Df    RSS Df Sum of Sq     F    Pr(>F)    
1    104 52.583                                 
2    103 31.156  1   21.4268 84.81 4.948e-15 ***
3    101 25.517  2    5.6391 11.16 4.178e-05 ***
---
```

We know that lower the RSS value, more the variability. As observed from the above figure, Model 3 has the lowest RSS value. This means model 3 has the highest variability. Also, the p-values are less than 0.05.

One of the ways to choose the best model is by comparing R^2. The variability of the model based on R^2 values is as below:

Model 1: 85.49%

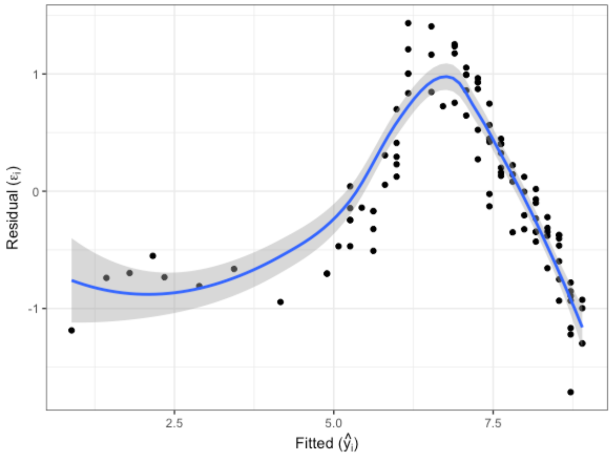
Model 2: 91.40%

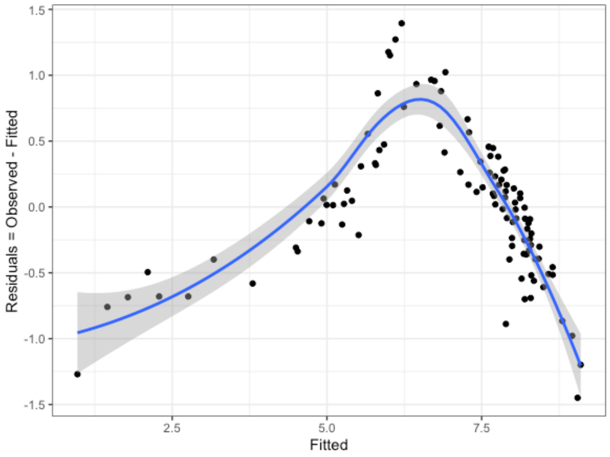
Model 3: 92.95%

Therefore, Model 3 is the best model.

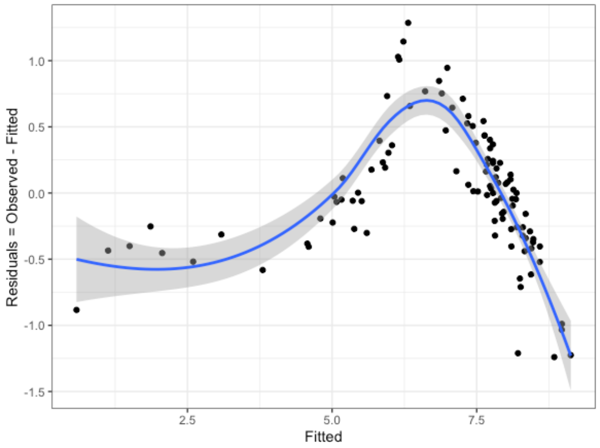
Section D - Interpretation (20%)

- The two plots given below of Model 1 and 2 do not satisfy the assumptions of the linear model as the residuals are not homegeneous, and they are not normally distributed.





However, the model 3 below (lm3) is much closer to satisfying the assumptions of linear regression than the previous models.



- Clock speed depends on many other factors and not just the time period, like the number of transistors, chores and power density. Therefore, Moore's law is not a good prediction for the advancement of computational power.