Which laptop should I buy?

(A Regression Analysis)

Amitangshu Dasgupta (161082)

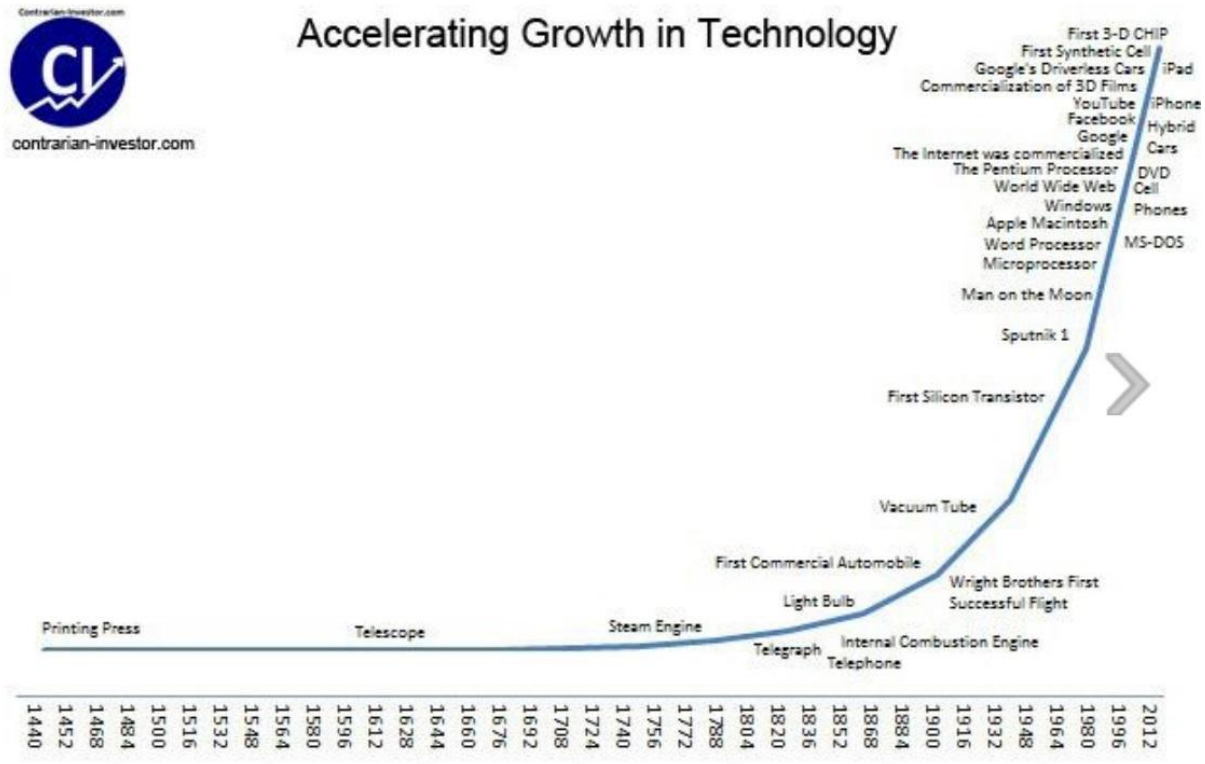
M.Sc. Statistics

IIT Kanpur

INTRODUCTION

This is an age of a technological revolution. Ever since the beginning of civilization, humans have innovated. And that's what set us apart from the other animals in the animal kingdom. We invented tools for hunting at first. And then a giant step was taken in around 3500 BC when we invented the wheel. Gradually, as we became civilized, our inventions became driven by convenience as compared to survival.

The years between 1760 and 1840 witnessed a new transition to manufacturing processes, otherwise known as the Industrial Revolution. Things were mass produced in factories and the common man got the opportunity to afford factory made stuff that used to be only a privilege for the rich until then. This trend went on with every new invention that we made as they were commercialized for the general public.



From: seekingalpha.com/article/453871-the-promise-of-accelerating-growth-in-technology

Between 1945 and 1947, Alan Turing, a British Mathematician, worked on the design of the ACE (Automatic Computing Engine) and presented a paper on 19 February 1946, which was the first detailed design of a stored-program computer. Computers became affordable for the general public in the 1970s due to the mass production of the microprocessor starting in 1971. IBM launched the first portable personal computer in 1984 which was evolved into a laptop by Compaq in 1988. Back then, laptops were freakishly costly and only the richest could afford it.

In the year 2015, 163 million laptops were sold worldwide. With such huge demands comes huge production and naturally a lot of choices to pick from. Today, a large number of manufacturing companies are developing their own line of laptops, each with improved feature than their previous. We all have been at some point in our lives where we had to pick a laptop from a number of choices and had to think twice about which model would actually be worth the money. What would be the best buy?

We have developed an algorithm to compare a given number of laptops to a large number of laptops in our database and give you the best model that you can buy given a budget. Let’s find out how.

DATA DESCRIPTION

We had a dataset containing the brand, features and specifications of 836 laptops and each of their prices. Each row contained complete information about a particular laptop. The fields were as follows:

Maximum Screen Resolution: real
Type of RAM (DDR2, SDRAM or DDRRAM): categorical
RAM capacity: real
Processor Speed: real
Manufacturer (Apple, HP, Lenovo etc.): nominal
Infrared: logical
Bluetooth: logical
Docking Station: logical
Port Replicator: logical
Fingerprint: logical
Subwoofer: logical
External Battery: logical
CDMA: logical
Operating System (MacOS, WindowsVista, WindowsXP etc): nominal
Warranty Period: real
Price: real (*target variable*)

DATA PREPROCESSING

Handling Missing Observations

In the beginning we noticed that we had 146 missing observations. The fields from which data was primarily missing were 'Maximum Screen Resolution' and the 'RAM Type' (69 and 63 missing cases respectively). The other fields with missing observations were 'RAM Capacity' (12 cases) and 'Processor Speed' (2 cases).

Instead of using the commonly used mean imputation technique (or mode in case of categorical variables), we took a more educated approach to the imputations.

STEP1:

We partitioned the datasets into two parts, viz. one with only the complete rows with no missing observations and the other with incomplete rows from each of which at least one field was missing.

STEP2:

For each incomplete row, we took subsets from the completed dataset partition in which every categorical field was a match with the corresponding non-missing fields from the incomplete row under consideration. For any incomplete row, there were three possibilities:

> CASE1: Subset matching from the complete dataset contained more than two rows
>
> In this case, the non-missing real-valued fields in the incomplete row were used to get the nearest observations from the complete and matched dataset, based on Euclidean distance.
> For real-valued missing fields, the average of the corresponding fields from the two nearest complete observations were imputed. For categorical missing fields, the corresponding field from the nearest complete observation was imputed.
>
> CASE2: Subset matching from the complete dataset contained exactly one row
>
> Missing fields were imputed straight away from the matched observation
>
> CASE3: Subset matching from the complete dataset was empty
>
> In this case again, the Euclidean distance was used to find the nearest observation (based only on real-valued non-missing fields) and the missing fields were imputed straight away from the nearest observation (ignoring the categorical fields).

Encoding Categorical Variables

The categorical variables were encoded into columns of dummy variables. For each variable the number of dummies used was one less than the number of distinct categories in that variable. For example:

RAM Type had three categories: DDR2, SDRAM, DDRRAM

Encoded into (0,0), (0,1), (1,0) where two dummy variables would be used. The absence of all dummy variables would represent the first category, presence of each dummy variable would represent either of the other two classes.

MODEL FITTING

The dataset was split randomly into training (90%) and testing sets (10%).

Initial Model

The first linear model was fitted based on the training set. All variables were included into the model. But the estimated coefficients for the variable OS.6 (the 7th category of Operating System) turned out to be NA which indicated the presence of severe multicollinearity.

Multicollinearity

We calculated VIF for each variable and recursively dropped variables with VIF more than 4 while fitting new models at every step. At the ninth step, we found the VIF of none of the remaining variables to be greater than 3.

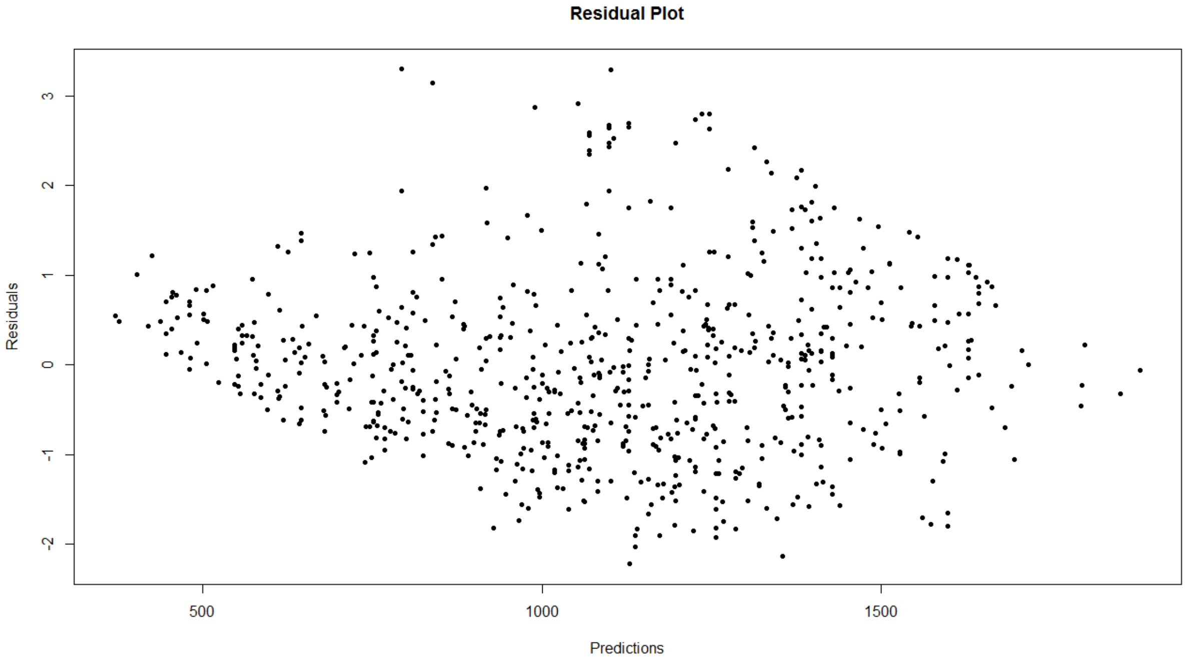
Model Selection

After dropping a number of regressors because of multicollinearity, we were still left with a large number of insignificant regressors. So we used the stepwise (forward and backward) model selection to get rid of insignificant regressors. We were left with only the regressors significant at 5% level of significance.

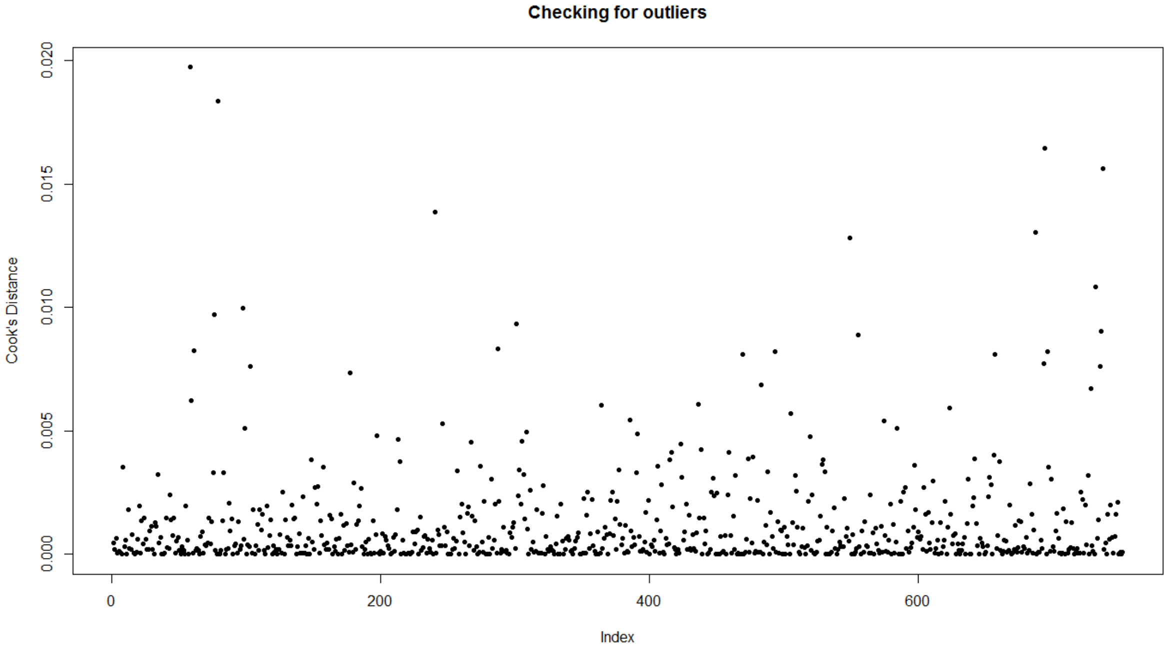
The model obtained finally gave a *residual standard error of 244.2* on 727 degrees of freedom with an *adjusted R squared value of 0.612*.

MODEL ADEQUECY

After we got our final model, we ran some model adequacy diagnostics. We used residual plot at first (plot of residuals vs. the predictions for all observations in the training set), which didn't seemed to display any kind of pattern.

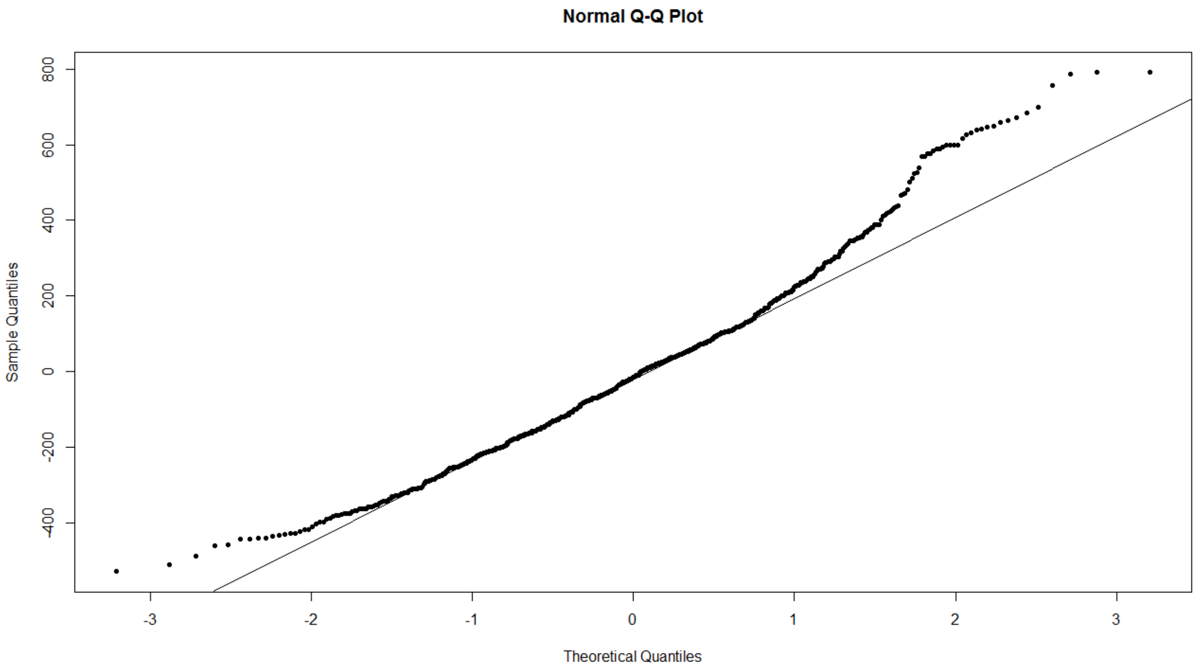


Checking for outliers in the dataset using cooks distance:

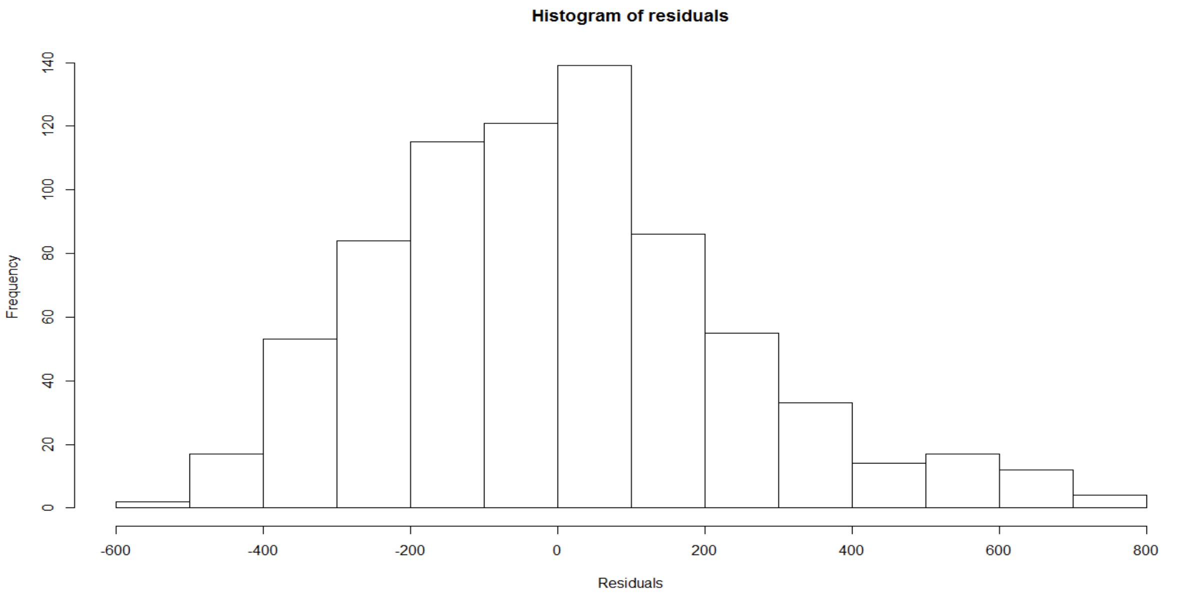


Clearly there are no outliers detected since the Cook's Distance for all observations in the training set is less than 0.02.

Checking for normality using qqplot:



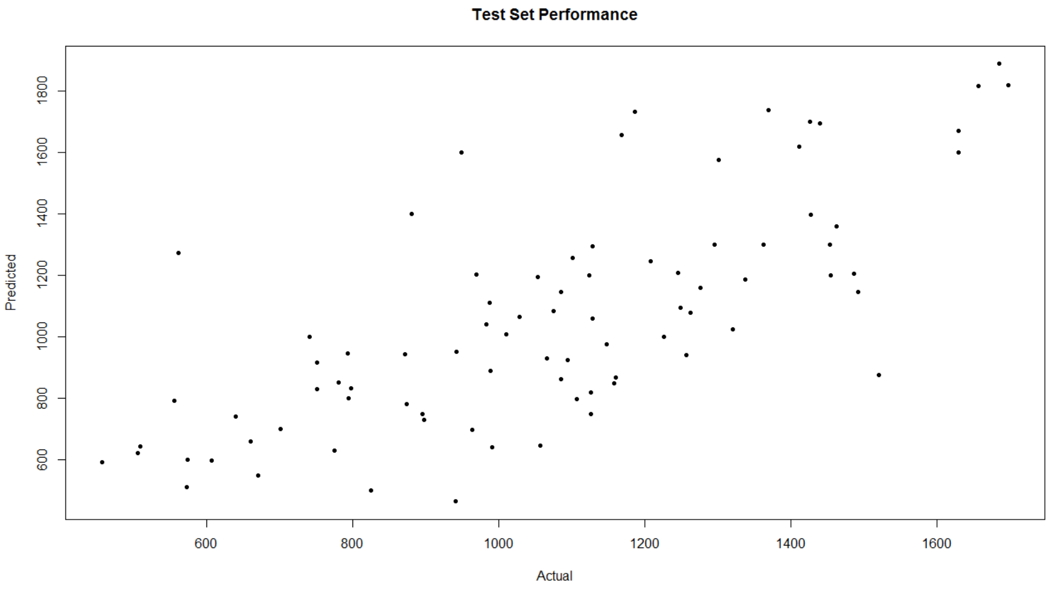
We don't seem to have a problem with the normality assumption of the residuals. Let's see how they are distributed.



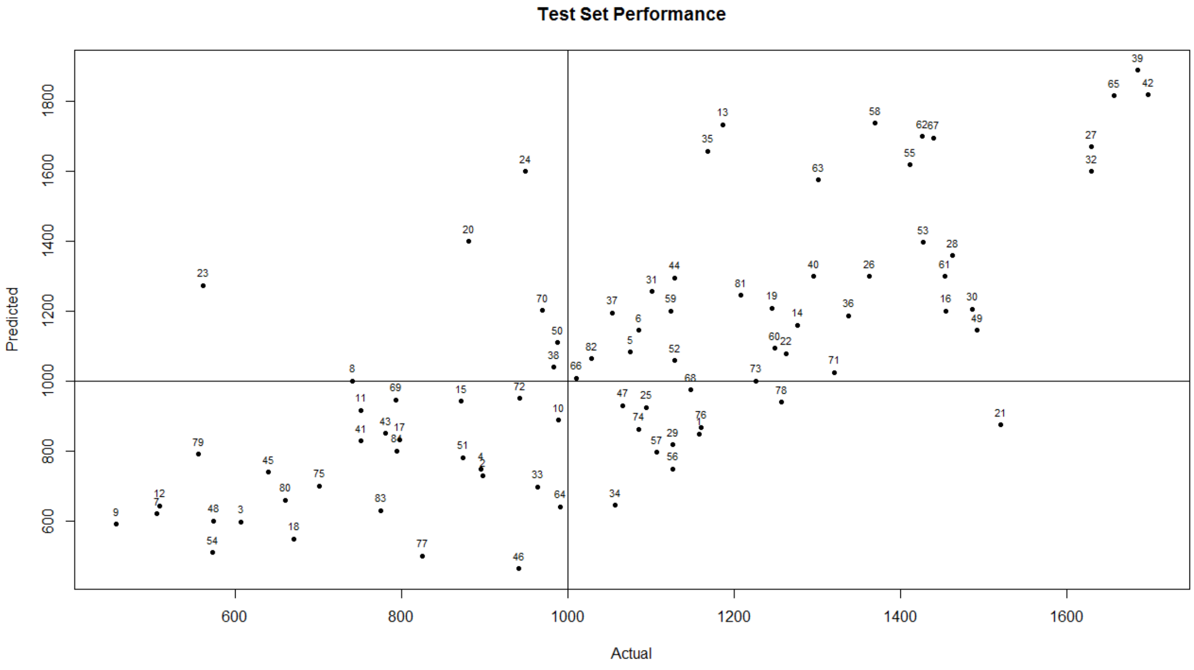
The distribution seems to be moderately symmetric about 0. So we can go ahead with our linear model.

CHOOSING THE BEST LAPTOP

Running the linear model on the test data, we obtained the following plot.



The plot shows in the y-axis, the predicted cost of the laptops in the test data based on its brand, features and specifications and comparing them with the price of other laptops data in our database. The x-axis shows the actual price. In other words, the y-axis shows how much a given laptop should cost and the x-axis shows how much it actually costs. By comparing the values, we can easily find the laptop that would be the best value for our money. For convenience, we added lines to the plot:



From the test set, the best value for money buy seems to be observation number 24.

Observation 24 is as follows:

Brand: TOSHIBA
RAM Type: DDR2
RAM Capacity: 512Mb
Processor: Intel Core Duo
Processor Speed: 1.8 GHz
Additional Features: Docking System
Operating System: WinXP Professional
Warranty: 3 years

We can use this model and of course a different dataset instead of the test set to find the cheapest laptop with the highest potential among a predetermined set of laptops. For example, we are confused whether to buy an Apple or a Dell laptop. We can feed in our choices to the model and obtain a plot as above and make our decisions likewise.