# Analysis of Differences Across Types of Interior Parts of Computer and Computer Price

Jingcheng Lu †, Ziwen Sun †, Xiaole Yu *,†

University of Toronto, Toronto, Canada

* Corresponding Author Email: 15000240215@xs.hnit.edu.cn

† These authors contributed equally.

**Abstract.** The enhancement of digital computers takes an active part in promoting the development of different aspects of the world. In the meantime, interior parts of computers are also rapidly enhanced and refined. Various combinations of inner parts of a computer would play a decisive factor in the cost so the aim of this article would discuss the relationship between the performance parameters of computers and the prices. This article divides the performance parameters of computers into nine aspects, such as speed, RAM size, etc., With an emphasis on the connection between the price of the computer and its performance parameters. This article employs a multiple linear regression and a logistic regression model to estimate the pricing using computer configuration parameters. The MSE, AIC and other parameters are established to measure the goodness of fit for models. In order to forecast prices for computers with certain performance specifications, an optimal model is finally developed.

**Keywords:** Computer price, performance parameter, linear regression.

## 1. Introduction

Computers are closely related to various industries related to life, such as transportation, medicine, water and electricity supply, and manufacturing, and are an indispensable tool for human beings today. In fact, as early as 1976, Brian and Gary proposed that the computer industry was one of the largest industries in the U.S, accounting for a huge market [1]. According to Statista, the computer penetration rate around the world increased year by year [2]. In the past two or three years, the emergence of virtual currencies such as Bitcoin has opened up a new market. The behavior of mining is very important for high-performance graphics cards. The demand for high-performance graphics cards has gradually increased, and the prices of many high-performance graphics cards have also fluctuated. In addition to the graphics card, the performance of other accessories of the computer is gradually improving. However, as computer technology has been updated, so has the price of computers, even under the influence of the epidemic, the computer industry will maintain a growth rate of more than 10% while occupying a market of more than $300+billion [3]. This chapter will link the price of a computer with the configuration of its parts, hoping to find a formula to define the relationship between its price and the performance of its accessories.

According to Grosch's analysis, a computer's speed significantly impacts its cost. There is Grosch's law, which states that the price of a computer is proportional to the square of its system performance. For example, when the performance of the system increases a hundredfold, the price increases tenfold [4]. However, in the current market environment, Grosch's law has been proved to be too general and useless [5]. At present, there is not enough research to define a very accurate relationship between the price of computers and the performance parameters of various configurations. Therefore, if a linear regression model of price and part parameters is defined, it is possible to understand the specific connection between computer prices and individual component parameters.

## 2. Methodology

This article will cover the following concepts:

### 2.1. Multiple Linear Regression

Simple linear regression does not apply when the variable is greater than 1. To predict response Y on multiple predictor variables $x_i$, an approach is needed that presents the relationship between response Y and variables $x_i$ in the form of

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \ldots + \beta_p x_p + \varepsilon \quad (1)$$

And the coefficient $\beta_i$ is estimated to be $\hat{\beta}_i$ and the response could be predicted by the formula of

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 \ldots + \hat{\beta}_p x_p \quad (2)$$

And the residual sum of squares (RSS) is defined as

$$RSS = \Sigma e_i^2 \quad (3)$$

Where $e_i$ is defined as

$$e_i = y_i - \hat{y}_i \quad (4)$$

The goodness of fit is measured by RSS. When the value of RSS is large, it means that the predicted response is far from the actual value, and the model doesn't fit very well [6].

### 2.2. AIC and MSE

Another two ways to measure the goodness of fit are Akaike Information Criterion (AIC) and MSE, which take the form of

$$AIC = n * \ln\left(\frac{SEE}{n}\right) + 2k \quad (5)$$

$$SSE = \Sigma(\hat{Y}_i - Y_1)^2 \quad (6)$$

$$MSE = SSE/n \quad (7)$$

Where k is the number of predictors, n is the sample size.

The smaller AIC, the smaller error and the models fits better [7]. And when the predicted response close to the true response the MSE will be small [8].

### 2.3. Correlation of two variables

In addition, a good linear regression model also needs to determine whether there is a correlation between variables. Correlation determines the strength of the relationship between the two variables, the closer its absolute value is to 1the stronger the correlation between the two variables. And r=Cor (X, Y), $R^2=r^2$ measures the fit of a model [9]. And Cor($x_i$, $x_j$) can also be used to check the correlation between $x_i$and $x_j$. The Pearson correlation coefficient takes the following formula

$$\frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\Sigma(x_i-\bar{x})^2}\sqrt{\Sigma(y_i-y)^2}} \quad (8)$$

### 2.4. Normalization

This is a method of data manipulation to remove the effects of order of magnitude differences on regression models and subsequent analysis. This method scales all data equally so that all data are in the same order of magnitude. The formula for this method is

$$x' = \frac{x-u}{\delta} \quad (9)$$

Where $u$ is the mean of the variable and $\delta$ is the variance of variables. The term $x'$ is called scaled *x* [10].

### 2.5. Basic computer data

The data comes from kaggle and shows the price of each computer and its corresponding performance parameters. Therefore, the data from Jonathan Bouchet five years ago serves as our sample, with a total of 6259 observations. This dataset collects data about the price and speed of the computer, hard drive (*HD*), presence or absence of *CD*, etc. The speed of the computer means how much data the connection can transfer. The hard drive is a storage device which can magnetically store and retrieve digital data. The hardware in a computing device can be referred to as *RAM* which saves contents as long as power is not turned off. The compact disc is a storage device that stores and records audio and video. Active Directory stores information and makes this information easily located and utilized by users. Different numbers of columns correspond to different computer data. The price of the data is defined as a dependent variable, others such as *speed*, *HD*, etc. are defined as independent variables, and there are a total of 9 dependent variables.

### 2.6. Data preprocessing

The second step is to preprocess the data. Firstly, the "yes" and "no" of the three dummy variables of *CD*, *multi* and *premium* in the independent variables are changed to 1 and 0 of numbers. Secondly, use the formula

$$x' = \frac{x-u}{\delta} \tag{10}$$

To standardize all variables. Thirdly, tabulate standardized data in r, which is a necessary prerequisite for subsequent statistical operations.

## 3. Results and Discussion

### 3.1. First assumption

The most basic linear model is employed, including a dependent variable of price and nine independent variables of *speed*, *HD*, *RAM*, s*creen*, *CD*, *multi*, *premium*, *ADS*, trend, and estimated independent variables and dependent variables present a kind of the simplest linear relationship

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} \ldots + \beta_p x_{ip} \tag{11}$$

The $i$ in the subscript refers to the $i_{th}$ computer, and the number n after the i in the subscript refers to the n-th component parameter. For example, $x_{32}$ refers to the parameter of the second component--hm of the third computer. which is 170. This assumption is initial and not optimized, and subsequent optimizations will operate on this model.

### 3.2. Result and analysis

Table I presents the results of the first linear regression model. In this model, the $R^2$ is 0.7756, which is not very high. The estimate coefficients correspond to B0-B9, and the P-value represents the significance of the coefficient which is displayed in the rightmost column Pr(>|t|). When the P-value is greater than 0.05, it means that the coefficient is insignificant and may have little effect on the dependent variable. The final model needs to satisfy the P-value of all variables less than 0.05, which indicates that they all will have a significant effect on the response. In the results of the first regression model, it is found that the P-value of intercept is 1, indicating that the intercept has less effect on Y, and this model is not perfect.

**Table 1.** The result for the first linear regression

| Coefficients | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| Intercept | $-2.6 \times 10^{-14}$ | $6.0\times 10^{-3}$ | 0.000 | 1 |
| Speed | $3.4\times 10^{-1}$ | $6.7\times 10^{-3}$ | 50.364 | < 0.001 *** |
| HD | $3.5\times 10^{-1}$ | $1.2\times 10^{-2}$ | 28.311 | < 0.001 *** |
| Ram | $4.7\times 10^{-1}$ | $1.0\times 10^{-2}$ | 45.265 | < 0.001 *** |
| Screen | $2.0\times 10^{-1}$ | $6.2\times 10^{-3}$ | 30.776 | < 0.001 *** |
| CD | $5.2\times 10^{-2}$ | $8.2\times 10^{-3}$ | 6.402 | < 0.001 *** |
| Multi | $6.2\times 10^{-2}$ | $6.8\times 10^{-3}$ | 9.141 | < 0.001 *** |
| Premium | $-2.6\times 10^{-1}$ | $6.3\times 10^{-3}$ | -41.259 | < 0.001 *** |
| ADS | $8.5\times 10^{-2}$ | $6.6\times 10^{-3}$ | 12.809 | < 0.001 *** |
| Trend | $-7.0\times 10^{-1}$ | $8.5\times 10^{-3}$ | -82.470 | < 0.001 *** |
| Multiple $R^2$: 0.776, Adjusted $R^2$: 0.776 | | | | |

Notes: ***represents 0.001 significance

### 3.3. Second assumption

In order to optimize this model, it is necessary to look at the correlation coefficient between the variables, which is presented in Table II. In this table, if the absolute value of correlation coefficient between the two variables is greater than 0.4, the two variables may be correlated. There are seven pairs of variables with correlation coefficients greater than 0.4, they are (*HD, RAM*), (*HD, CD*), (*RAM, CD*), (*multi, CD*), (*trend, speed*), (*trend, HD*), (*trend, CD*).

**Table 2.** The Pearson correlation coefficient of variables

| | Speed | HD | Ram | Screen | CD | Multi | Premium | ADS | Trend |
|---|---|---|---|---|---|---|---|---|---|
| Speed | 1.000 | 0.372 | 0.235 | 0.189 | 0.258 | 0.0842 | 0.114 | -0.215 | 0.405 |
| HD | 0.372 | 1.000 | 0.778 | 0.233 | 0.504 | 0.093 | 0.197 | -0.323 | 0.578 |
| Ram | 0.235 | 0.777 | 1.000 | 0.209 | 0.439 | 0.045 | 0.197 | -0.182 | 0.277 |
| Screen | 0.189 | 0.233 | 0.209 | 1.000 | 0.129 | -0.002 | 0.0188 | -0.094 | 0.189 |
| CD | 0.258 | 0.503 | 0.439 | 0.129 | 1.000 | 0.432 | 0.216 | -0.061 | 0.446 |
| Multi | 0.084 | 0.092 | 0.045 | -0.002 | 0.432 | 1.000 | 0.125 | -0.030 | 0.211 |
| Premium | 0.114 | 0.197 | 0.197 | 0.018 | 0.216 | 0.125 | 1.000 | -0.152 | 0.042 |
| ADS | -0.215 | 0.323 | -0.182 | -0.094 | -0.061 | -0.030 | -0.152 | 1.000 | -0.319 |
| Trend | 0.405 | 0.577 | 0.277 | 0.189 | 0.446 | 0.211 | 0.0421 | 0.319 | 1.000 |

Note: Table II shows the correlation coefficients of pairs between variables.

### 3.4. Second linear model

A second linear regression model is built by adding correlated variables to the first model. The relevant data for the second linear model are shown in Table III. It can be found that in this model, there are still many variables that are not significant, but $R^2$ is 0.8266, which is higher than the first model, which shows that it may be effective to add correlated variables as new variables to the first model.

**Table 3.** The result for the second linear regression

| Coefficients | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| Intercept | 0.173 | 0.007 | 24.199 | < 0.001 *** |
| Speed | 0.362 | 0.006 | 58.575 | < 0.001 *** |
| HD | 0.466 | 0.011 | 40.829 | < 0.001 *** |
| Ram | 0.453 | 0.010 | 43.249 | < 0.001 *** |
| Screen | 0.192 | 0.006 | 34.665 | < 0.001 *** |
| CD | 0.054 | 0.007 | 7.238 | < 0.001 *** |
| Multi | 0.052 | 0.006 | 8.485 | < 0.001 *** |
| Premium | -0.273 | 0.006 | -48.885 | < 0.001 *** |
| ADS | -0.067 | 0.007 | -9.453 | < 0.001 *** |
| Trend | -0.734 | 0.008 | -88.483 | < 0.001 *** |
| HD×Ram | 0.020 | 0.006 | 3.326 | < 0.001 *** |
| HD×CD | -0.058 | 0.012 | -5.003 | < 0.001 *** |
| Ram×CD | 0.001 | 0.010 | 0.060 | 0.952453 |
| CD×Multi | NA | NA | NA | NA |
| Speed×Trend | -0.105 | 0.006 | -16.414 | < 0.001 *** |
| HD×Trend | -0.170 | 0.008 | -20.362 | < 0.001 *** |
| CD×Trend | -0.042 | 0.008 | -5.025 | < 0.001 *** |
| Multiple $R^2$: 0.827, Adjusted $R^2$: 0.826 | | | | |

Notes: ***represents 0.001 significance.

### 3.5. StepAIC

Some variables in the second model are not significant, try to remove some variables and observe the AIC of the entire model, and repeat this step until the smallest AIC and its corresponding model are found - this is also the final optimization of the second model version, this is the first three models. This step can be implemented in r with stepAIC. This code can remove a variable in order to reduce AIC and repeat this operation until AIC cannot be reduced. The last step of stepAIC can obtain a final model that cannot be re-optimized. This is the third model mentioned above. The results of the last step of stepAIC are shown in Table IV.

**Table 4.** The final step of stepAIC

| Coefficients | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| Intercept | 0.173 | 0.007 | 24.546 | < 0.001 *** |
| Speed | 0.362 | 0.006 | 58.693 | < 0.001 *** |
| HD | 0.466 | 0.011 | 41.147 | < 0.001 *** |
| Ram | 0.453 | 0.010 | 44.158 | < 0.001 *** |
| Screen | 0.192 | 0.005 | 34.701 | < 0.001 *** |
| CD | 0.054 | 0.007 | 7.267 | < 0.001 *** |
| Multi | 0.052 | 0.006 | 8.498 | < 0.001 *** |
| Premium | -0.273 | 0.006 | -48.904 | < 0.001 *** |
| ADS | -0.067 | 0.007 | -9.456 | < 0.001 *** |
| Trend | -0.734 | 0.008 | -88.711 | < 0.001 *** |
| HD×Ram | 0.020 | 0.006 | 3.494 | < 0.001 *** |
| HD×CD | -0.058 | 0.010 | -5.877 | < 0.001 *** |
| Speed×Trend | -0.105 | 0.006 | -16.428 | < 0.001 *** |
| HD×Trend | -0.170 | 0.008 | -20.368 | < 0.001 *** |
| CD×Trend | -0.042 | 0.008 | -5.261 | < 0.001 *** |
| Multiple $R^2$: 0.827, Adjusted $R^2$: 0.826 | | | | |

Note: The stepAIC code removes some potentially irrelevant variables and retains those relevant variables, so that the predicted model can fit better. In the end, it can be found that the p-values of all

variables are less than 0.05, indicating that all variables in the third model can have an impact on the response y. ***represents 0.001 significance.

### 3.6. Third linear model

The model obtained in the last step of stepAIC is defined as the third model, and the following Table V is some data of the third model. In the third model, all variables are significant and the $R^2$ is 0.8266, which is the same as the $R^2$ of the second model. However, the AIC of the third model is 6824.396 and the AIC of the second model is 6826.793. so the third model will outperform the second model and be its final optimized version.

**Table 5.** The result of the third model

| Coefficients | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| Intercept | 0.173 | 0.007 | 24.546 | < 0.001 *** |
| Speed | 0.362 | 0.006 | 58.693 | < 0.001 *** |
| HD | 0.466 | 0.011 | 41.147 | < 0.001 *** |
| Ram | 0.453 | 0.010 | 44.158 | < 0.001 *** |
| Screen | 0.192 | 0.005 | 34.701 | < 0.001 *** |
| CD | 0.054 | 0.007 | 7.267 | < 0.001 *** |
| Multi | 0.052 | 0.006 | 8.498 | < 0.001 *** |
| Premium | -0.273 | 0.006 | -48.904 | < 0.001 *** |
| ADS | -0.067 | 0.007 | -9.456 | < 0.001 *** |
| Trend | -0.734 | 0.008 | -88.711 | < 0.001 *** |
| HD×Ram | 0.020 | 0.006 | 3.494 | < 0.001 *** |
| HD×CD | -0.058 | 0.010 | -5.877 | < 0.001 *** |
| Speed×Trend | -0.105 | 0.006 | -16.428 | < 0.001 *** |
| HD×Trend | -0.170 | 0.008 | -20.368 | < 0.001 *** |
| CD×Trend | -0.042 | 0.008 | -5.261 | < 0.001 *** |
| Multiple $R^2$: 0.827, Adjusted $R^2$: 0.826. | | | | |

Notes: ***represents 0.001 significance.

### 3.7. Comparison of linear model 1 and model 3

The third model is an optimized version of the second model, it is only necessary to compare the third model with the first model. The effect of comparing fitting is shown in Table VI. The smaller the RSS, the better the model fits, obviously the third linear model is better than the first one.

**Table 6.** Anova of model 1 and model 3

| | Res.Df | RSS | Df | Sum of Sq | F | Pr(>F) |
|---|---|---|---|---|---|---|
| L1 | 6249 | 1404.5 | - | - | - | - |
| L3 | 6244 | 1084.8 | 5 | 319.66 | 367.98 | <0.001 *** |

Note: Anova is a method in R which could present the RSS of two linear model. In Table VI, L1 refers to the first linear model and L3 refers to the third model.

### 3.8. Final model

The third linear model serves as the final model to predict the price. More information is attached in Table VII.

**Table 7. More information about the third model**

| Coefficients | Estimate | Pr(>\|t\|) |
|---|---|---|
| Intercept | 0.173 | < 0.001 *** |
| Speed | 0.362 | < 0.001 *** |
| HD | 0.466 | < 0.001 *** |
| Ram | 0.453 | < 0.001 *** |
| Screen | 0.192 | < 0.001 *** |
| CD | 0.054 | < 0.001 *** |
| Multi | 0.052 | < 0.001 *** |
| Premium | -0.273 | < 0.001 *** |
| ADS | -0.067 | < 0.001 *** |
| Trend | -0.734 | < 0.001 *** |
| HD×Ram | 0.020 | < 0.001 *** |
| HD×CD | -0.058 | < 0.001 *** |
| Speed×Trend | -0.105 | < 0.001 *** |
| HD×Trend | -0.170 | < 0.001 *** |
| CD×Trend | -0.042 | < 0.001 *** |
| RMSE: 0.417, $R^2$: 0.826, MAE: 0.317 | | |

Note: RMSE is the root of MSE, so the MSE of this model is 0.17413971. ***represents 0.001 significance.

Note that Y and $x_i$ in this model do not represent the component parameters of the price and the computer itself, but are standardized numbers, so there is a problem about the prediction of the price of a given computer's own component parameters?

### 3.9. Prediction

The following demonstrates the method of price prediction. Firstly, normalize each variable in $x_1$ to $x_9$ with the formula $x' = \frac{x-u}{\delta}$ and obtain the scaled variable $x_i'$. Secondly, put all the scaled variables in model 3 and the term scaled response $y_i'$ is obtained. Thirdly, reverse the scale formula to remove scale of the response $y_i'$ with the new formula $y = y_i' * \delta + u$. The y is the final prediction of price.

## 4. Conclusion

This article attempts to establish an optimization model for predicting computer prices based on a set of datasets about computer component parameters including *speed*, *HD*, *RAM*, *screen*, *CD*, *multi*, *premium*, *ADS* and *trend*. $R^2$, AIC and other indicators are used in R to the fit of the measure model is good or bad. After comparison, it is found that the fit of the linear regression model containing correlated variables is better than the simplest linear regression. Then, according to the principle that the smaller the AIC, the better, the complex linear regression is optimized, the unnecessary variables are removed, and the final model is obtained. The model produces an $R^2$ score of 82.66% and a MSE of 0.174. The above results indicate that the components of the computer may influence each other, and their mutual influence may further affect the pricing of the computer. And 82.66% of the price can be explained by both individual variables--*speed*, *HD*, *RAM*, *screen*, *CD*, *multi*, *premium*, *ADS* and *trend* and correlated variables-- (*HD*, *RAM*), (*HD*, *CD*), (*RAM*, *CD*), (*multi*, *CD*), (*trend*, *speed*), (*trend*, *HD*), (*trend*, *CD*). And the price of the computer can be predicted from a set of given data of component parameters using the above method. In a more in-depth study, you can try to add more correlated variables or try to simulate a nonlinear model and compare the goodness of fits to choose a model to make a more accurate prediction of computer prices.

## References

[1] R. Brian and F. Gary, "A Study of Prices and Market Shares in the Computer Mainframe Industry," The Journal of Business, University of Chicago Press, vol. 49(2), 1976, pp. 194-218.

[2] T. Alsop, "Share of households with a computer at home worldwide from 2005 to 2019. Retrieved from https://www.statista.com/statistics/748551/worldwide-households-with-computer/.2022.

[3] Globe Newswire, "Mobile Computer Market Worth $3.3 Billion by 2029 - Exclusive Report by Meticulous Research". Retrieved from https://www.globenewswire.com/newsrelease/2022/07/06/2475220/0/en/Mobile-Computer-Market-Worth-3-3-Billion-by-2029-Exclusive-Report-by-Meticulous-Research.html. 2022.

[4] H.R.J. Grosch, "High Speed Arithmetic: The Digital Computer as a Research Tool," Journal of the Optical Society of America. 43(4), 1953, pp. 306–310.

[5] E. G. Cale, L. L. Gremillion and J. L. McKenney, "Price/Performance Patterns of U. S. Computer Systems," Communications of the ACM, vol. 22(4), 1979, pp. 225–233.

[6] J. Frost, Mean Square Error (MSE). Statistics By Jim. Retrieved from https://statisticsbyjim.com/regression/mean-squared-error-mse/.2022.

[7] H. Akaike, "A new look at the statistical model identification," IEEE Transactions on Automatic Control, 19 (6), 1974, pp. 716-723.

[8] G. James, D. Witten, T. Hastie and R. Tibshirani, An introduction to statistical learning (2nd ed.). 2013, pp. 29-37.

[9] G. Casella, Statistical inference (Second ed.). Pacific Grove, Calif.: Duxbury/Thomson Learning. 2002, pp. 556-557.

[10] E. Kreyszig, Advanced Engineering Mathematics (Fourth ed.). Wiley. 1979, pp. 880-881.