Data Analysis 2 & Coding 1 Final Term Project

Laptop Prices

Ersan Kucukoglu

Introduction

In this project, I'll discuss my results on how laptop specs drive laptop prices.The goal of this project is to explore how *laptop prices* differ based on their *processor* type controlled with some z (control/conditioning) variables.The confounding variable that selected into the different models are the *ram, operating system, and laptop types like a gaming,notebook, etc.* Mainly, I wanted to analyze the changes made on the effects of the technical features that change the prices of the laptops, which are so widely used today, with statistical methods. My main hypothesis is that elevated processors, such as the Intel Core i-3,5,7, and 9, are increasing laptop prices. I expect the lapto prices high in the case of the processor is i7, ram size is higher. It is one of the most important specifications, because the processor, also known as the CPU, provides the instructions and processing power the computer needs to do its work. The more powerful and updated your processor, the faster your computer can complete its tasks. By getting a more powerful processor, you can help your computer work faster. Briefly, the Core "i" names are primarily "high level" categorizations that help differentiate processors within a given generation. Core designations refer to relative improvements within a specific generation of processors. As the Core number increases, so do the capabilities of the processors, including higher core counts, faster clock speeds, more cache, and the ability to handle more RAM.

Data

For the analysis, I have used the Laptop Prices Data from *Kaggle*. The original dataset consists of 1303 observations which means 1303 different laptop and 13 variables with their specifications and prices.

The description of the *variables* are as follows:

1. laptop_ID - Numeric - Laptop ID Number
2. Company- String - Laptop Manufacturer
3. Product - String -Brand and Model
4. TypeName - String -Type (Notebook, Ultrabook, Gaming, etc.)
5. Inches - Numeric- Screen Size
6. ScreenResolution - String- Screen Resolution
7. Cpu - String -Central Processing Unit (CPU)
8. Ram - String- Laptop RAM
9. Memory - String- Hard Disk / SSD Memory
10. GPU -S tring- Graphics Processing Units (GPU)
11. OpSys - String- Operating System
12. Weight -S tring- Laptop Weight
13. Price_euros - Numeric- Price (Euro)

In order to get useful data for further analysis and to make it ready to use in statistical models, the data requires some data cleaning and transformations for some variables.

- *Data Cleaning*

The first step in data cleaning process is to make the variables ready which are planned to used in the statistical model.In order to filter laptops with Intel processors, firstly, by separating CPU variable, i3, i5, i7 values were extracted and converted the variable into a numeric variable. The same process was done for the Ram variable. Unnecessary columns (laptop_ID, ScreenResolution, Gpu, Weight, Inches) were dropped which were not planned to used in the analysis. Following first cleaning, the data was left with 1086 observations and 8 variables, but it still required transformations to be prepared for regressions. After exploration data and running the distribution of the price, intelCore_i(CPU), RamGB, Inches (Screen Size), OpSys (Operating System), TypeName (gamin, ultrabook, notebook, etc.) variables, I observed the most frequent values of the potential confounding variables which are RamGB, OpSys, and TypeName to create dummy variables for them. The next process was to creating binary for our confounding variables so I could use them as dummy variables in our statistical model.I created dummy variables for the OpSys and TypeName categorical variables. Also, I created dummy variables for the values of the intelCore_i variable, because the number of processors are discrete. I have processor types with the number of 3, 5, 7. These are not continuous variables, but discrete variables. According to association_figs1, in the Appendix, when I checked its association with the price variable on the scatter plot , it is almost linearly. It is not curved twist shape, but still it has three distinct values (3,5,7). In this case, I should think about transformation of this variable. Instead of using intelCore_i as it is, I should create dummy variable for each i3,i5 and i7.In addition, after checking the distribution of the price data, prices variable was converted in log to normal distribute the right hand side distribution. After seeing the result in the association_figs2, see in the Appendix, I decided to take log on price since the data was right skewed and I had to normalize the distribution.

- *Descriptive Statistics*

Table 1: Descriptive statistics

	Mean	Median	SD	Min	Max	P05	P95
Price (€)	1237.56	1096.08	663.20	339.00	6099.00	468.00	2499.00
ln_Price (€)	6.99	7.00	0.51	5.83	8.72	6.15	7.82
Ram (GB)	9.04	8.00	5.14	4.00	64.00	4.00	16.00
i3	0.13	0.00	0.33	0	1	0.00	1.00
i5	0.39	0.00	0.49	0	1	0.00	1.00
i7	0.49	0.00	0.50	0	1	0.00	1.00
OpSys_W10_d	0.84	1.00	0.37	0	1	0.00	1.00
OpSys_W7_d	0.04	0.00	0.20	0	1	0.00	0.00
OpSys_Linux_d	0.05	0.00	0.21	0	1	0.00	0.00
OpSys_macOS	0.01	0.00	0.10	0	1	0.00	0.00
OpSys_No_OS	0.05	0.00	0.22	0	1	0.00	1.00
Gaming Type	0.18	0.00	0.39	0	1	0.00	1.00
Notebook Type	0.53	1.00	0.50	0	1	0.00	1.00
Ultrobook Type	0.17	0.00	0.37	0	1	0.00	1.00

Based on descriptive statistics for the sample data. The table indicates that 49% of the laptops in the sample data have i7 intelcore processor, 39% of them are i5, 13% of them are i3. The distribution of the price is right skewed with an average of 1237.56 and median of 1096.08. In addition, it can be inferred from the table that the most of the laptops (84%) have Windows 10 Operating System and the almost half of the

laptops (53%) are Notebook type. Based on the association_figs1, in the Appendix, the association between the lnprice and intelCore_i variables is linearly positive. It also has positive correlation between lnprice and Ram. To get an overview of how the variables are associated with one another,correlation matrix is created to extract the correlation coefficient for each of them. The correlation matrix shows the level of association of price with our dependent and confounding variables. From the correlation matrix, it can be observed that there is positive association of lnprice with intelCore_i_7 with the correlation coefficient of 0.56. On the other hand, the price was negatively associated with the other two processors i3 and i5 with the correlation coefficient -0.53 and -0.20. Also, there is a strong positive correlation between lnprice and Ram with the correlatio coefficient of 0.70. One interesting find that this matrix indicated was that there was no as such correlation between the the OpSys(Linux,macOS,NoOS,Win10,Win7) variables and the lnprice, and theirs correlation coefficients are less than 0.2. The correlation matrix can be found in Appendix.

Models

I run 4 different regression models and one by one adding the confounding variable. For comparing the regression results, the Summary of the regressions model can be found in the Appendix (Table2). I emphasized that the expected relationship with processors is hypothesized that the price of laptops with i7 processors will be higher compared to i5 and i3. The reason is that the i7 processor makes computers stronger and faster, Intel's most powerful chips, Core i7 CPUs tend to be a lot more expensive. Out of the many reasons one of the most important one would be that the market for i7 isn't as big as compared to its mid range counterpart - the i3 and i5. Latter are used more often because computer with basic and medium processing power has a larger customer base than that of the more powerful (gaming/graphic oriented) PCs. So, lesser the demand, higher the manufacturing cost due to smaller mass production and thereby higher price per unit.

1. *lnprice - intelCore_i*

$$lnprice = \beta_0 + \beta_1 intelCorei3 + \beta_2 intelCorei5$$

The first regression model, which is log - level model, since I had an expected association that the pricing of laptops with i7 processors would be higher, I used it as the basis for the statistical model. I regressed ln prices on the intel processor dummy variables and found that if the laptop has intel core i7 processor, the intercept tells us the ln price is going to be 7.28. If the laptop has i3 processor, the ln price is 1.021 less than 7.28 on average. If the laptop has i5 processor, then the ln price is going to be 0.428 less than 7.28 on average.

The second hypothesis was that laptops with higher RAM size would be more expensive. For instance, if you are a gamer or designer then you have to have better equipment, you have to maintain the temperature more even, the space needs to be clean. Making a 16GB RAM chip means building twice as many gates as a 8GB chip, in the same space. These requires tighter tolerances, so higher costs, so higher prices.

2. *lnprice - intelCore_i + RamGB*

$$lnprice = \beta_0 + \beta_1 intelCorei3 + \beta_2 intelCorei5 + \beta_3 RamGB$$

The second model accounts for the first confounding variable which is the Ram of the laptop. The beta coefficient of this model states that keeping other factors constant the ln price of laptops is 0.044 higher for each higher ram size on an average. The confidence interval for the RamGB variable is 99% respectively.

Third, laptop type (such as gamer, notebook) and operating system have an impact on laptop price. Due to the high demand created by the powerfulness of gamer laptops, their prices are expected to be quite high.

3. *lnprice - intelCore_i + RamGB + OpSys*

$$lnprice = \beta_0 + \beta_1 intelCorei3 + \beta_2 intelCorei5 + \beta_3 RamGB + \beta_4 OpSysLinux + \beta_5 OpSysMacOS + \beta_6 OpSysWin10 + \beta_7 Op$$

The third model accounts for the second confounding variable which is the Operating System of the Laptop. Here I have taken the No OS dummy varible as the base and the run the statistical model. The beta coefficient for the Linux OS that keeping everything else constant,in comparison to No OS,the prices of the Linux OS Laptops tend to be 0.067 lower on an average with zero level of confidence. The beta coefficient for the Mac OS in comparison to No OS,the prices of the Mac OS Laptops tend to be 0.66 higher on an average with 99% significance level. Win10 OS in comparison to No OS, the ln prices of the Win10 OS Laptops tend to be 0.27 higher on average with the 99% significance level. Win7 OS in comparison to No OS, the ln prices of the Win7 OS Laptops tend to be 0.61 higher on average with the 99% significance level.

My final hypothesis is that laptops without an operating system should be cheaper than those with an operating system. Also, due to the majority of users of the windows operating system, laptops with windows are also expected to be cheaper compared to macOS.

4. *lnprice - intelCore_i + RamGB + OpSys + TypeName*

$$lnprice = \beta_0 + \beta_1 intelCorei3 + \beta_2 intelCorei5 + \beta_3 RamGB + \beta_4 OpSysLinux + \beta_5 OpSysMacOS + \beta_6 OpSysWin10 + \beta_7 Op$$

The last model in our analysis is adding the Laptop TypeName dummy variables as confounding variable. TypeName_Gaming dummy variable is selected as the base since I assume that the gaming laptops tend to be higher price than others. The beta coefficient for the Notebook laptops in comparison to Gamer Laptops,the prices of the Notebook laptops tend to be 0.26 lower on an average with 99% significance level. Also, the Ultrabook laptops in comparison the gamer laptops, its prices tend to be 0.05 higher on an average with 90% significance level.

Conclusion

In sum, i observed how our regression models verified some of the hypotheses and rejected others. The model 1 validated our hypothesis that laptops with Intel Core i7 processors were more expensive than those with other CPUs. Furthermore, when comparing model 1 and model 2, i have seen that adding the Ram control variable was advantageous as the R squared increased, and the variable's coefficient is significant at the 99 percent confidence interval.

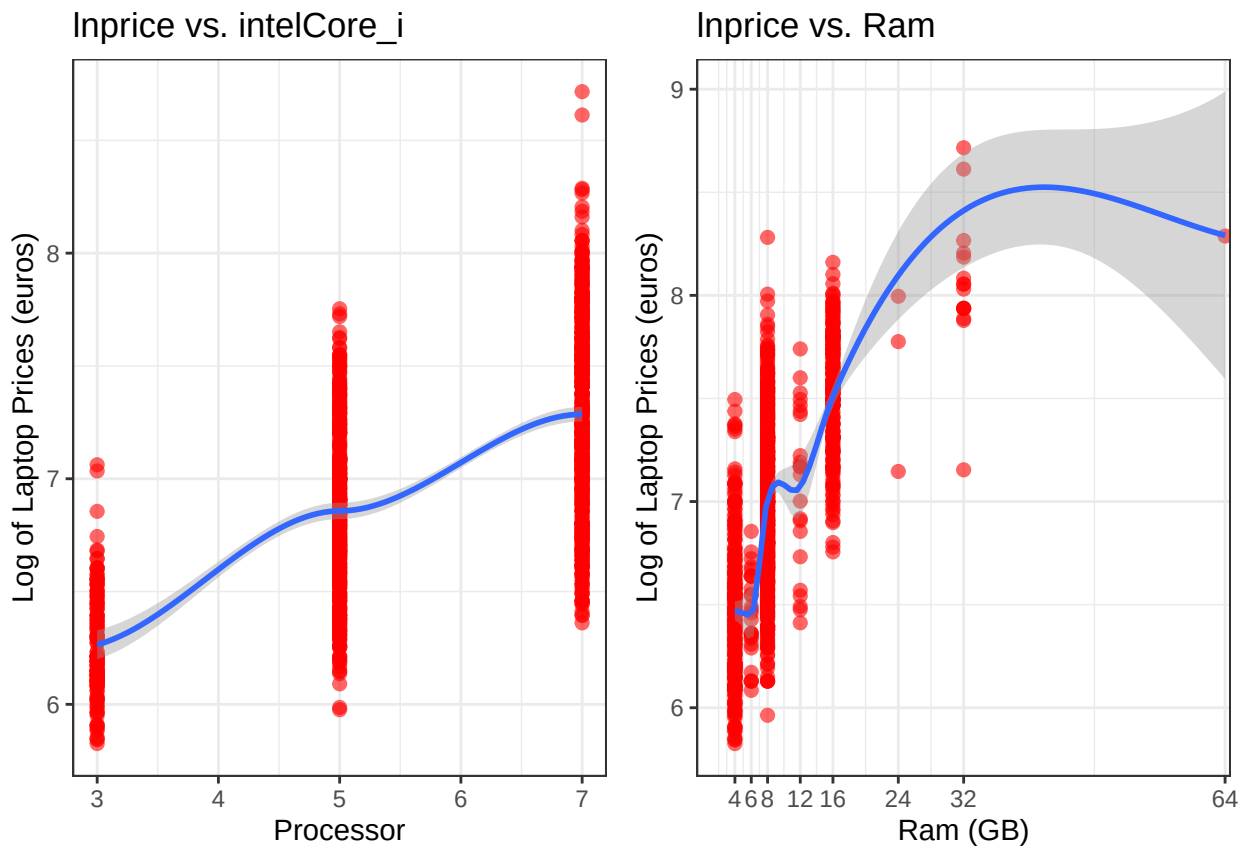
With the addition of the OpSys dummy variable, i was able to conclude from model 3 that the price of No OS Laptops will be relatively lower than other laptops with macOS, Win10, or Win7 at a % confidence interval. However, because the zero confidence interval for the Linux OS dummy variable does not verify the third hypothesis about the laptops' operating system.

In the last model, it can validate that gamer laptops are more expensive than Notebook type laptops, and it is significant at 99% significance level. However, the summary regression table shows that gamer laptops are cheaper than Ultrabook type of laptops, but it is not that significant.

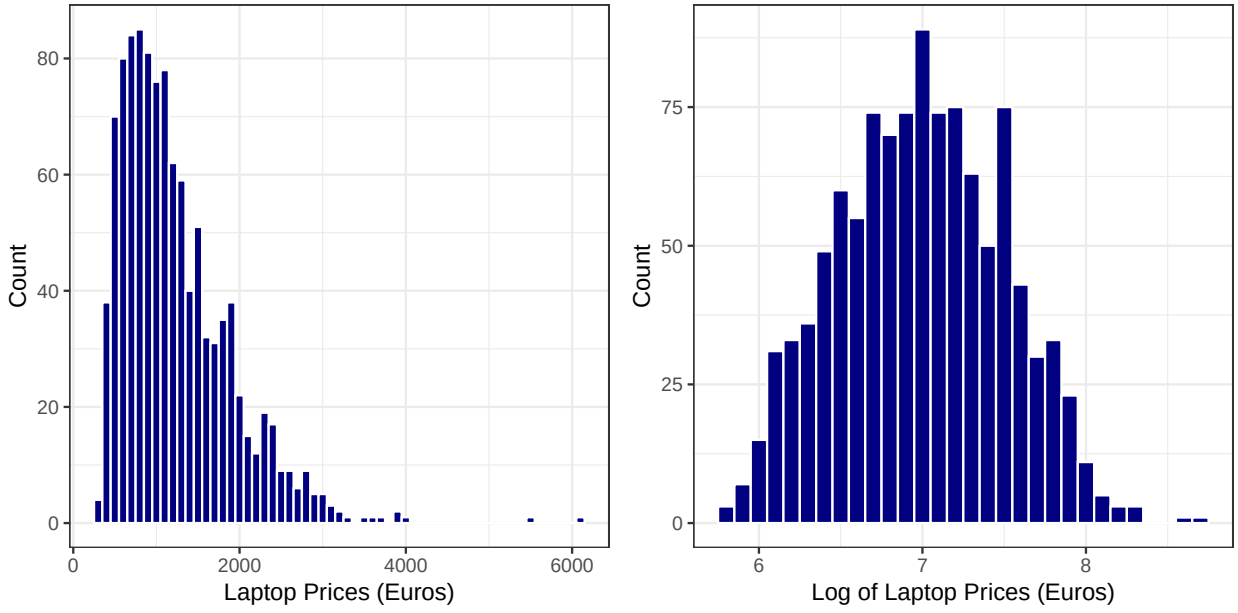
My preferred model is the 4th model. It is selected as a preferred model with higher adjusted and predicted R-squared values. The adjusted R squared increases only if the additional term improves the model more than would be expected by chance, and it can potentially drop when the predictors are of low quality. The preferred model, according to the Summary of Regression Table, is the reg4. The reason for choosing this model is that the Adjusted R squared increased to 0.71, which improved the R squared, implying a better quality predictor.

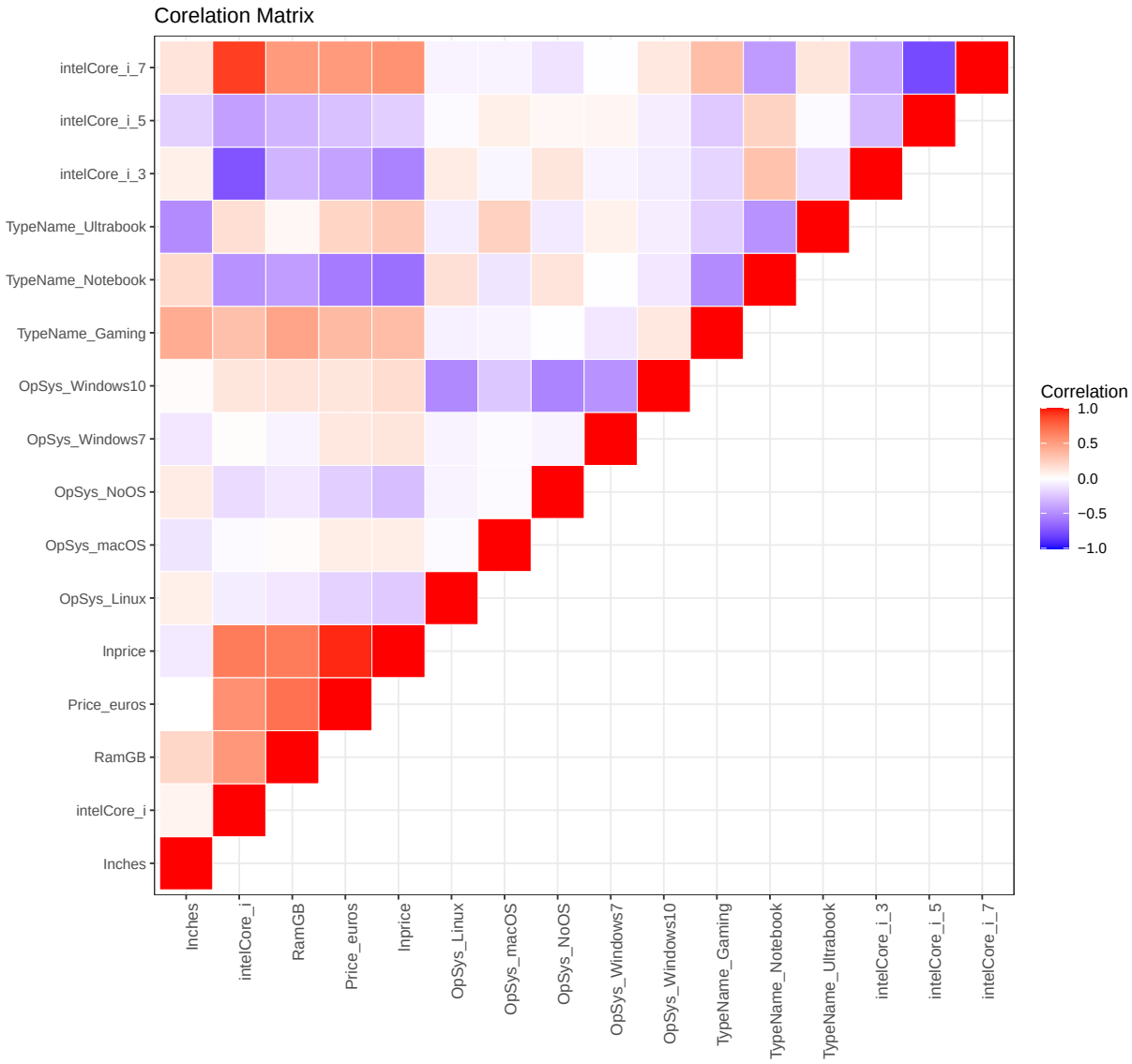
Appendix

Association between the ln price and processor, Ram



The Distribution of the prices and log prices




Corelation Matrix
intelCore_i_7
intelCore_i_5
intelCore_i_3
TypeName_Ultrabook
TypeName_Notebook
TypeName_Gaming
OpSys_Windows10
OpSys_Windows7
OpSys_NoOS
OpSys_macOS
OpSys_Linux
lnprice
Price_euros
RamGB
intelCore_i
Inches
Inches
intelCore_i
RamGB
Price_euros
lnprice
OpSys_Linux
OpSys_macOS
OpSys_NoOS
OpSys_Windows7
OpSys_Windows10
TypeName_Gaming
TypeName_Notebook
TypeName_Ultrabook
intelCore_i_3
intelCore_i_5
intelCore_i_7
Correlation
1.0
0.5
0.0
−0.5
−1.0

[H]

Table 2: Regression Summary Table

	reg1	reg2	reg3	reg4
(Intercept)	7.286***	6.765***	6.509***	6.985***
	(0.0184)	(0.0486)	(0.0603)	(0.0704)
intelCore_i_3	-1.022***	-0.7110***	-0.6606***	-0.5155***
	(0.0277)	(0.0371)	(0.0345)	(0.0304)
intelCore_i_5	-0.4285***	-0.2147***	-0.2164***	-0.1417***
	(0.0255)	(0.0285)	(0.0270)	(0.0238)
RamGB		0.0440***	0.0430***	0.0373***
		(0.0040)	(0.0039)	(0.0035)
OpSys_Linux			-0.0679	-0.3427***
			(0.0466)	(0.0564)
OpSys_macOS			0.6609***	0.1266
			(0.0748)	(0.0793)
OpSys_Windows10			0.2753***	-0.0689
			(0.0408)	(0.0506)
OpSys_Windows7			0.6157***	0.2533***
			(0.0585)	(0.0605)
OpSys_NoOS				-0.4145***
				(0.0567)
TypeName_Notebook				-0.2620***
				(0.0243)
TypeName_Ultrabook				0.0593*
				(0.0266)
Observations	1,086	1,086	1,086	1,086
R2	0.44401	0.58343	0.64508	0.71495