Laptop Price Prediction

Qingyin Ge
Citina Liang
Tianle Zhu
Yuezhi Liao
Yunuo Ma
Yiren Wang

Summary

1. Raw Data, 11 Variables
2. Clean Data, Split Data
3. Transform Y Variable
4. Model Selection
5. Model Diagnostic & Remedial measures :
    a. Multicollinearity
    b. Non Normality
    c. Non Constant Variance
    d. Outliers
6. Conclusion

Raw Data

- Company Name
- Product Name
- Laptop Type
- Screen Inches
- Screen Resolution
- CPU Model
- Random Access Memory (RAM)
- Memory
- Gpu Characteristic
- Operating System
- Laptop's Weight
- Laptop's Price (Response variable)

Company <chr>	Product <chr>	TypeName <chr>	Inches <dbl>	ScreenResolution <chr>	Cpu <chr>	Ram <chr>	Memory <chr>
Apple	MacBook Pro	Ultrabook	13.3	IPS Panel Retina Display 2560x1600	Intel Core i5 2.3GHz	8GB	128GB SSD
Apple	Macbook Air	Ultrabook	13.3	1440x900	Intel Core i5 1.8GHz	8GB	128GB Flash Storage
HP	250 G6	Notebook	15.6	Full HD 1920x1080	Intel Core i5 7200U 2.5GHz	8GB	256GB SSD
Apple	MacBook Pro	Ultrabook	15.4	IPS Panel Retina Display 2880x1800	Intel Core i7 2.7GHz	16GB	512GB SSD
Apple	MacBook Pro	Ultrabook	13.3	IPS Panel Retina Display 2560x1600	Intel Core i5 3.1GHz	8GB	256GB SSD

Cpu <chr>	Ram <chr>	Memory <chr>	Gpu <chr>	OpSys <chr>	Weight <chr>	Price_euros <dbl>
Intel Core i5 2.3GHz	8GB	128GB SSD	Intel Iris Plus Graphics 640	macOS	1.37kg	1339.69
Intel Core i5 1.8GHz	8GB	128GB Flash Storage	Intel HD Graphics 6000	macOS	1.34kg	898.94
Intel Core i5 7200U 2.5GHz	8GB	256GB SSD	Intel HD Graphics 620	No OS	1.86kg	575.00
Intel Core i7 2.7GHz	16GB	512GB SSD	AMD Radeon Pro 455	macOS	1.83kg	2537.45
Intel Core i5 3.1GHz	8GB	256GB SSD	Intel Iris Plus Graphics 650	macOS	1.37kg	1803.60

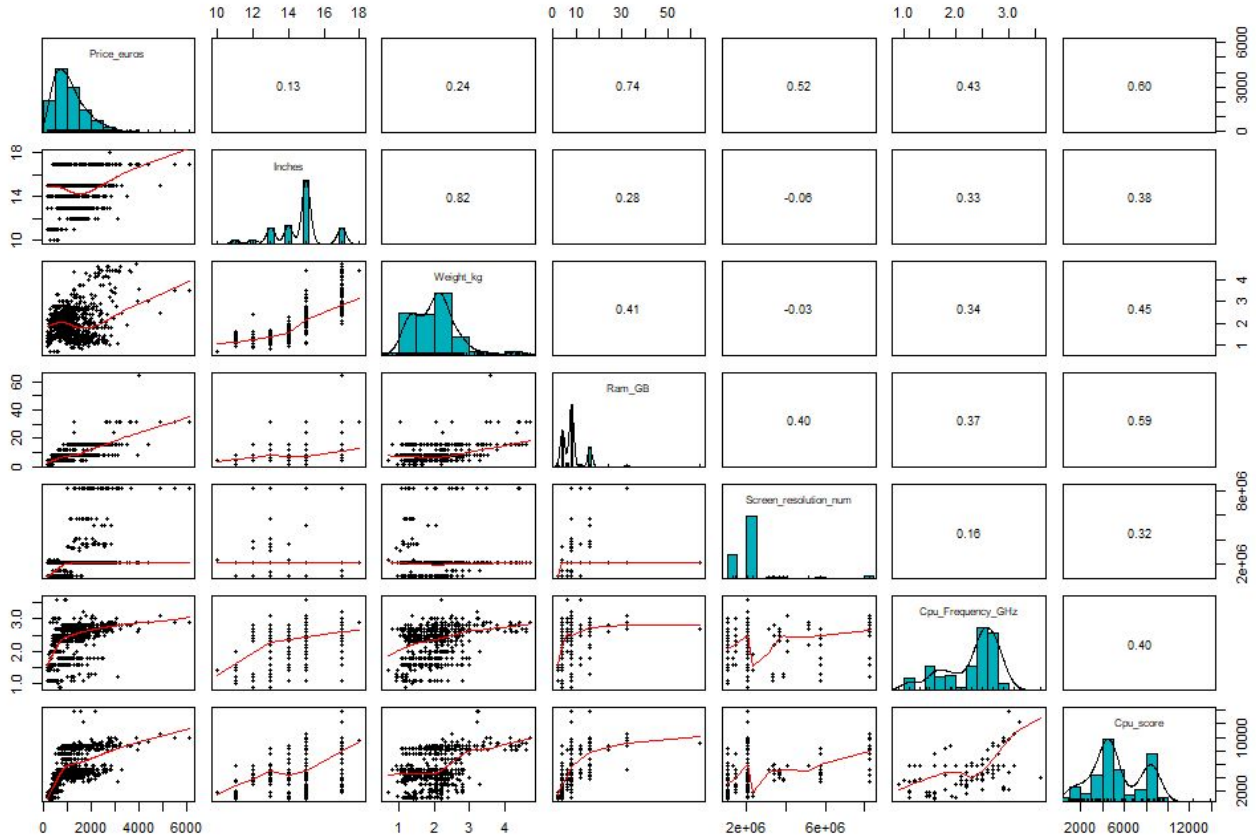
Clean Data

- **Product and Type Name:**
  - Dropped
- **Screen Resolution:**
  - Separated into screen resolution and touchscreen (0, 1)
- **CPU and GPU:**
  - Replaced numerically by CPU/GPU benchmarks
- **Memory:**
  - Separated into four categories (Flash, HDD, Hybrid, SSD)
- **Company/Operational Systems:**
  - changed to 10/ 8 qualitative predictor variable
- **Inches:**
  - Change the size of screen (inches) to integer
- **Weight :**
  - Change Weight(char) variable into Weight_kg(dbl)
- **Ram:**
  - Keep only the size of ram with unit GB.

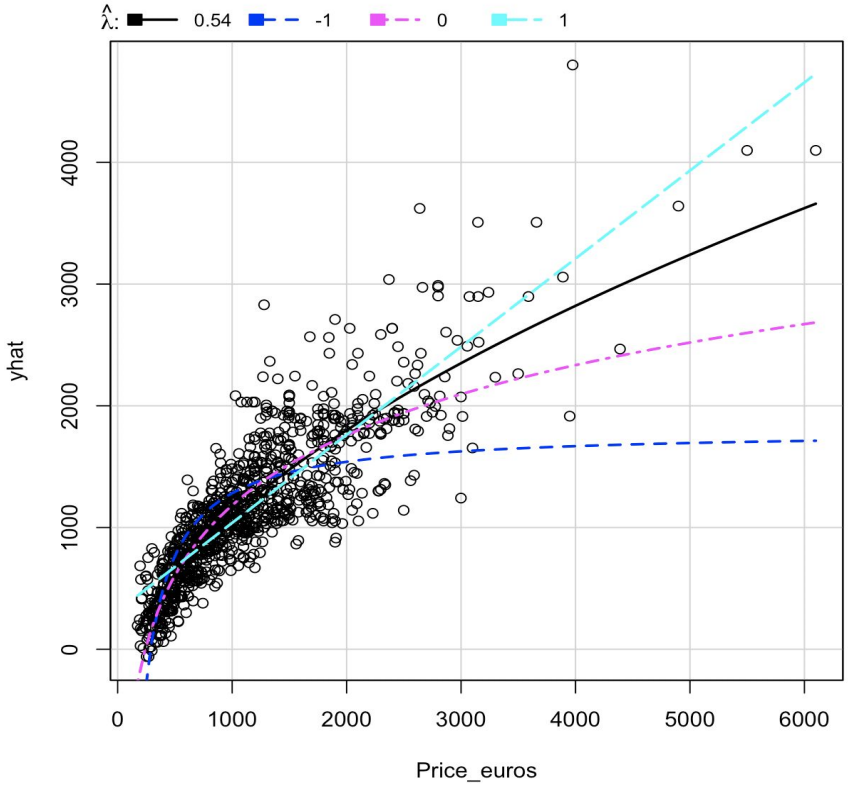
Inches	Price_euros	Weight_kg	Ram_GB	Touchscreen	Screen_resoluti…	Cpu_Frequency_G…	Cpu_score	Flash	HDD	Hybrid	SSD	Gpu_score
<int>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
13	1340.	1.37	8	0	4096000	2.3	5816.	0	0	0	128	2.1
13	899.	1.34	8	0	1296000	1.8	5816.	128	0	0	0	0.46
15	575	1.86	8	0	2073600	2.5	5816.	0	0	0	256	0.65
15	2537.	1.83	16	0	5184000	2.7	5816.	0	0	0	512	6.05
13	1804.	1.37	8	0	4096000	3.1	5816.	0	0	0	256	2.78

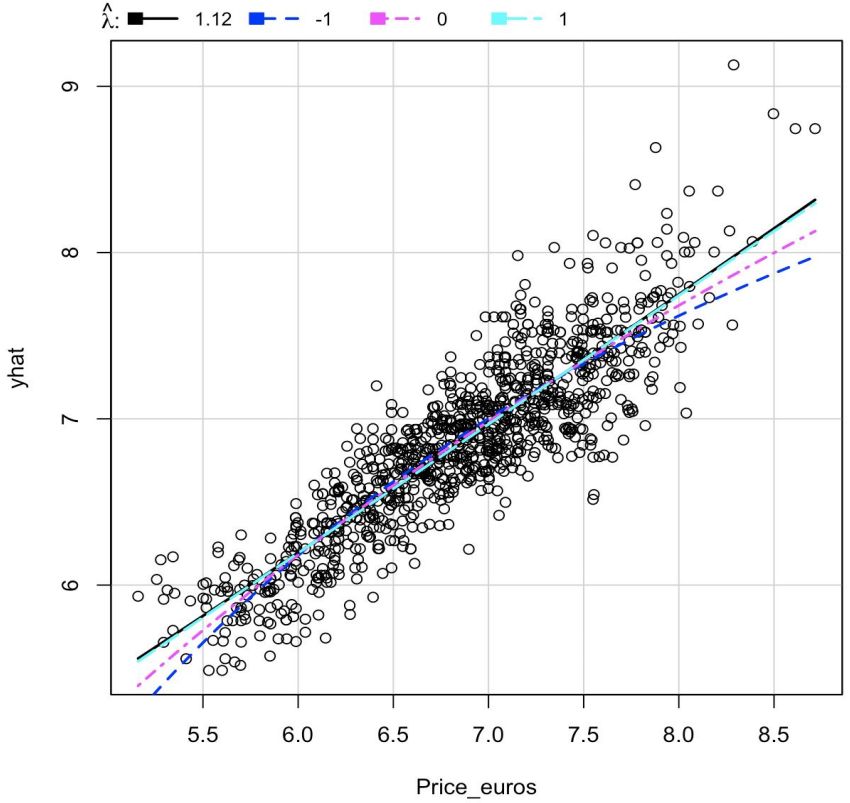
Acer	Asus	Dell	Fujitsu	HP	Huawei	Lenovo	MSI	Toshiba	Android	`Chrome OS`	Linux	`Mac OS X`	macOS	`No OS`	`Windows 10`
<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0
0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0

The Scatter Plot on Training Data



Transform Y to log(Y)





Model selection

Min_model: regress response variable on constant, no predictor variable

Full_model: regress response variable on all predictor variables

Use stepwise regression with forward and backward direction to find our final selected model, which contains:

Cpu_score, Ram_GB, Weight_kg, Cpu_Frequency_GHz, Windows 7, Screen_resolution_num, No OS, Acer, Toshiba, Touchscreen, Linux, Asus, macOS, Mac OS X, HP, Inches, SSD and Android.

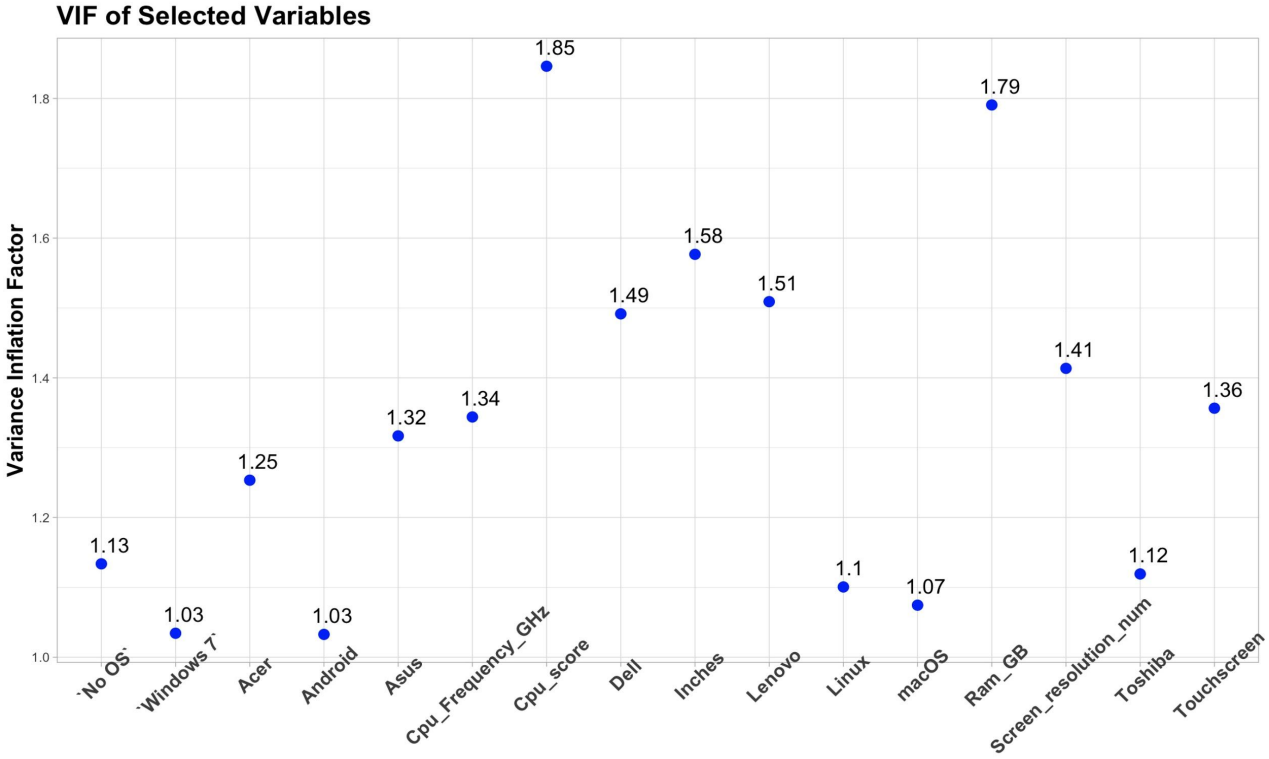
After fitting test data, we find out that the predicted R-square improved from 0.7317687 to 0.7786004.

```
Residuals:
     Min       1Q   Median       3Q      Max
-0.97248 -0.18896 -0.01379  0.17560  1.02487

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           5.744e+00  1.472e-01  39.026  < 2e-16 ***
Cpu_score             1.146e-04  5.146e-06  22.267  < 2e-16 ***
Ram_GB                3.893e-02  2.433e-03  15.999  < 2e-16 ***
Weight_kg            -1.042e-01  2.638e-02  -3.952 8.30e-05 ***
Cpu_Frequency_GHz     2.815e-01  2.099e-02  13.411  < 2e-16 ***
`Windows 7`           4.856e-01  5.006e-02   9.700  < 2e-16 ***
Screen_resolution_num 6.125e-08  7.937e-09   7.716 2.87e-14 ***
`No OS`              -2.702e-01  4.271e-02  -6.326 3.76e-10 ***
Acer                 -1.836e-01  3.585e-02  -5.121 3.63e-07 ***
Toshiba               2.812e-01  5.078e-02   5.537 3.93e-08 ***
Touchscreen           1.365e-01  3.083e-02   4.428 1.06e-05 ***
Linux                -1.655e-01  4.701e-02  -3.520  0.00045 ***
Asus                 -6.896e-02  3.041e-02  -2.268  0.02354 *
macOS                 2.711e-01  8.780e-02   3.087  0.00207 **
`Mac OS X`            2.820e-01  1.240e-01   2.273  0.02322 *
HP                    5.300e-02  2.427e-02   2.184  0.02922 *
Inches               -2.740e-02  1.234e-02  -2.220  0.02665 *
SSD                  -9.571e-05  4.800e-05  -1.994  0.04643 *
Android              -3.984e-01  2.110e-01  -1.888  0.05926 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2932 on 1010 degrees of freedom
Multiple R-squared:   0.78,     Adjusted R-squared:  0.7761
F-statistic:   199 on 18 and 1010 DF,  p-value: < 2.2e-16
```

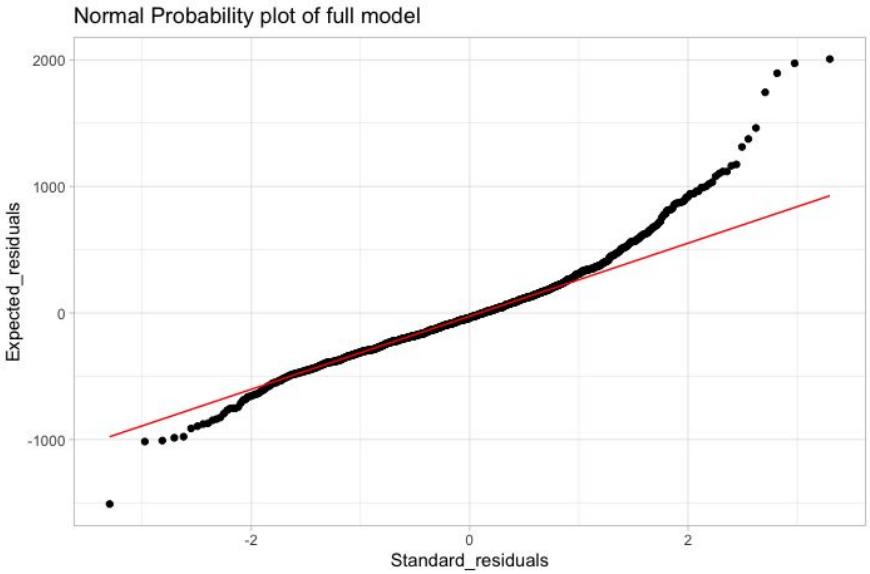
Model Diagnostic & Remedial measures - Multicollinearity



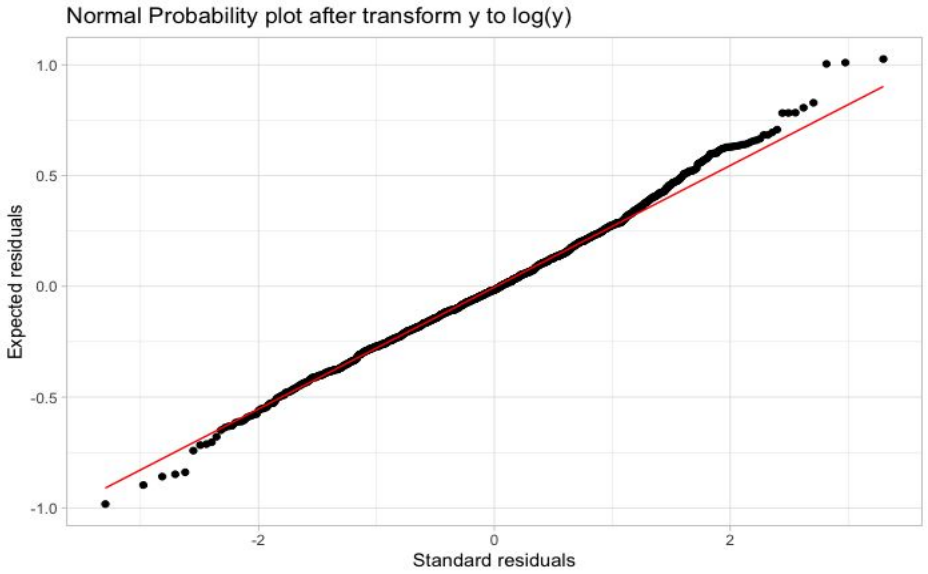
According to the graph, we don't have multicollinearity issue here.

Model Diagnostic & Remedial measures - Non Normality

Full Model

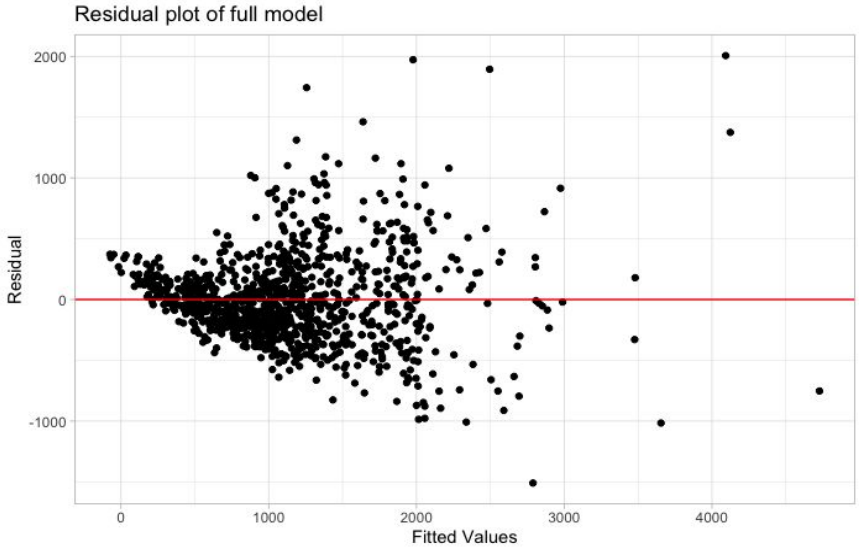


Transform Y



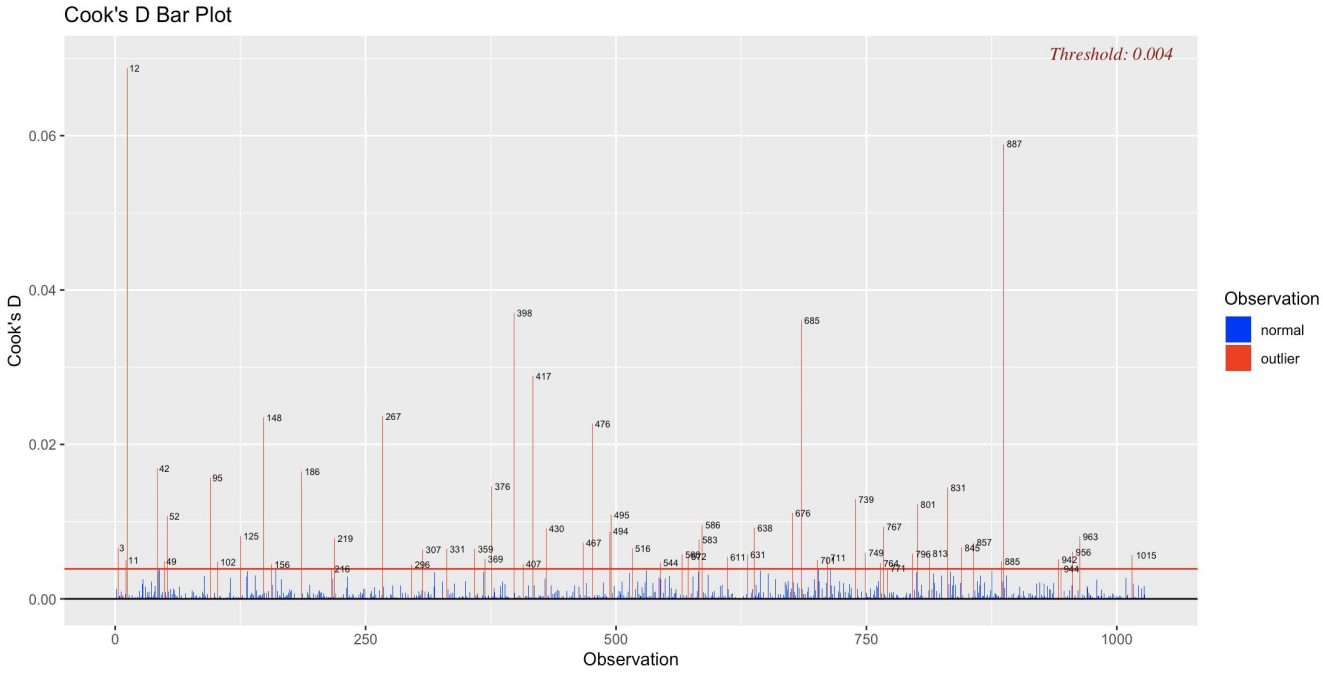
From the above Q-Q plot, we fixed the non-normality by transforming Y to log(Y).

Model Diagnostic & Remedial measures - Non Constant Variance



From the above plot, we can see that full model's residuals variance is not constant, so we decide to use weighted least square model. We calculate weight by using absolute residuals versus selected predictors. The right hand side graph shows we have almost constant variance now.
The predicted R square of the weighted model: 0.778793

Model Diagnostic & Remedial measures - Outliers



The Cook's Distance Bar Plot suggests that we have outliers in our data, so we use robust regression model to fix this problem. We use the same weight as the WLS model.
The result shows the predicted R square is 0.7790693

Conclusion

Our final model formula:

- **Price_euros** = exp{5.51+ 1.11e-04 * **Cpu_score** + 0.045 * **Ram_GB** - 0.155 * **Weight_kg** + 0. 253* **Cpu_Frequency_GHz** + 0.461 * **Windows_7** + 6.35e-08 * **Screen_resolution_num** - 0.261 * **No_OS** - 0.182 * **Acer** + 0.295 * **Toshiba** + 0.118 * **Touchscreen** - 0.156 * **Linux** - 0.0715 * **Asus** + 0.265* **macOS** + 0.26 * **Mac_OS_X** + 0.033 * **HP** - 2.69e-03 * **Inches** - 4.055e-05 * **SSD** - 0.331 * **Android**}

Example :

Inches <int>	Price_euros <dbl>	Weight_kg <dbl>	Ram_... <dbl>	Touchscreen <dbl>	Screen_resolution_num <dbl>	Cpu_Frequency_GHz <dbl>	Cpu_score <dbl>
13	1803.6	1.37	8	0	4096000	3.1	5815.869

- Y= exp(5.502 + 1.11e-04 * 5815.869 + 0.045 * 8 - 0.155* 1.37 + 0.253* 3.1 + 6.354e-08 * 4096000 + 0.265 *1 - 2.687e-03 * 13 - 4.055e-05 * 256) = 1918.8

Study Limitation:

- few data, dealing with outliers, variable selection...

Future Enhancement:

- Study on the coefficients that are not significant, include all/no categorical
- Try to get more data, filter out data to have less outliers, non linear model...

	Price_euros		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	5.50	5.22 – 5.79	**<0.001**
Cpu_score	0.00	0.00 – 0.00	**<0.001**
Ram_GB	0.05	0.04 – 0.05	**<0.001**
Weight_kg	-0.15	-0.21 – -0.10	**<0.001**
Cpu_Frequency_GHz	0.25	0.22 – 0.29	**<0.001**
Windows_7	0.46	0.38 – 0.54	**<0.001**
Screen_resolution_num	0.00	0.00 – 0.00	**<0.001**
No_OS	-0.26	-0.32 – -0.20	**<0.001**
Acer	-0.18	-0.24 – -0.13	**<0.001**
Toshiba	0.30	0.22 – 0.37	**<0.001**
Touchscreen	0.12	0.05 – 0.19	**0.001**
Linux	-0.16	-0.21 – -0.10	**<0.001**
Asus	-0.07	-0.13 – -0.01	**0.015**
macOS	0.26	0.18 – 0.35	**<0.001**
Mac_OS_X	0.27	0.18 – 0.35	**<0.001**
HP	0.03	-0.02 – 0.08	0.197
Inches	-0.00	-0.03 – 0.02	0.828
SSD	-0.00	-0.00 – 0.00	0.295
Android	-0.33	-0.85 – 0.18	0.208
Observations	1029		

References

- https://www.kaggle.com/ionaskel/laptop-prices
- https://www.cpubenchmark.net/high_end_cpus.html
- https://www.cpubenchmark.net/low_end_cpus.html
- https://www.cpubenchmark.net/midlow_range_cpus.html
- https://www.cpubenchmark.net/mid_range_cpus.html
- https://gpu.userbenchmark.com/


ngiyabonga
danke
謝謝
teşekkür ederim
спасибо
dank je
thank you
gracias
tapadh leat
mochchakkeram
bedankt
hvala
mauruuru
dziękuję
go raibh maith agat
sagolun
sukriya
kop khun krap
arigatō
takk
dakujem
obrigado
grazie
мерси
terima kasih
merci
감사합니다
ευχαριστώ