資四甲 110716035 吳宜昌

機器學習HW1\_Regression Analysis

程式碼語言：Python

資料集：HW1\_CarPrice

內容：

1. **資料清理與視覺化圖表（程式碼Part1. 資料清理與視覺化圖表）：**

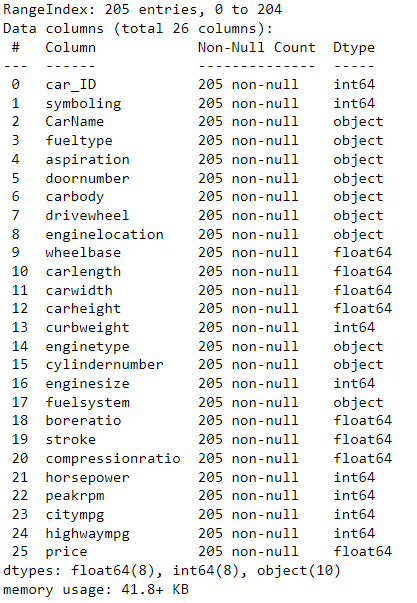


圖1 檢查是否有Null值

如上圖1，此資料並無任何Null值。

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

圖2資料集箱型圖

如上圖2，此資料有些許離群值，但是因為此資料為carPrice，代表這些離群值皆是現實生活中會遇到的真實情況，與本次要預測的車子價格有一定的相關性，且資料集只有205筆，因此選擇保留這些離群值或許會是比較的選擇。

1. **敘述性統計分析（程式碼Part2. 敘述性統計分析）：**

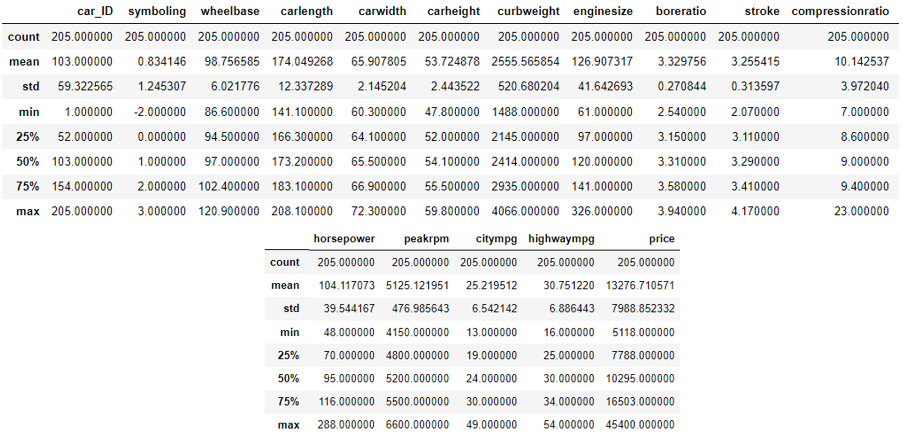


圖3 敘述性統計

如上圖3，此為資料集carPrice的敘述性統計，包含所有類別的平均數、標準差、最小值、4分之1位數、中位數、4分之3位數、最大值。

1. **特徵相關性分析（程式碼Part3. 特徵相關性分析）：**



圖4 熱力圖

如上圖4，此為資料集carPrice的特徵相關性分析。可從此圖得知與Price高度正相關的有carwidth、horsepower、curbweight、enginesize（由低到高排列）。

1. **資料分割與建置迴歸模型（程式碼Part4. 資料分割與建置迴歸模型）：**

因為與Price高度正相關最大的為enginesize，因此本次以enginesize為自變數，Price為應變數建立迴歸模型，且資料分割以7：3的比例切割。

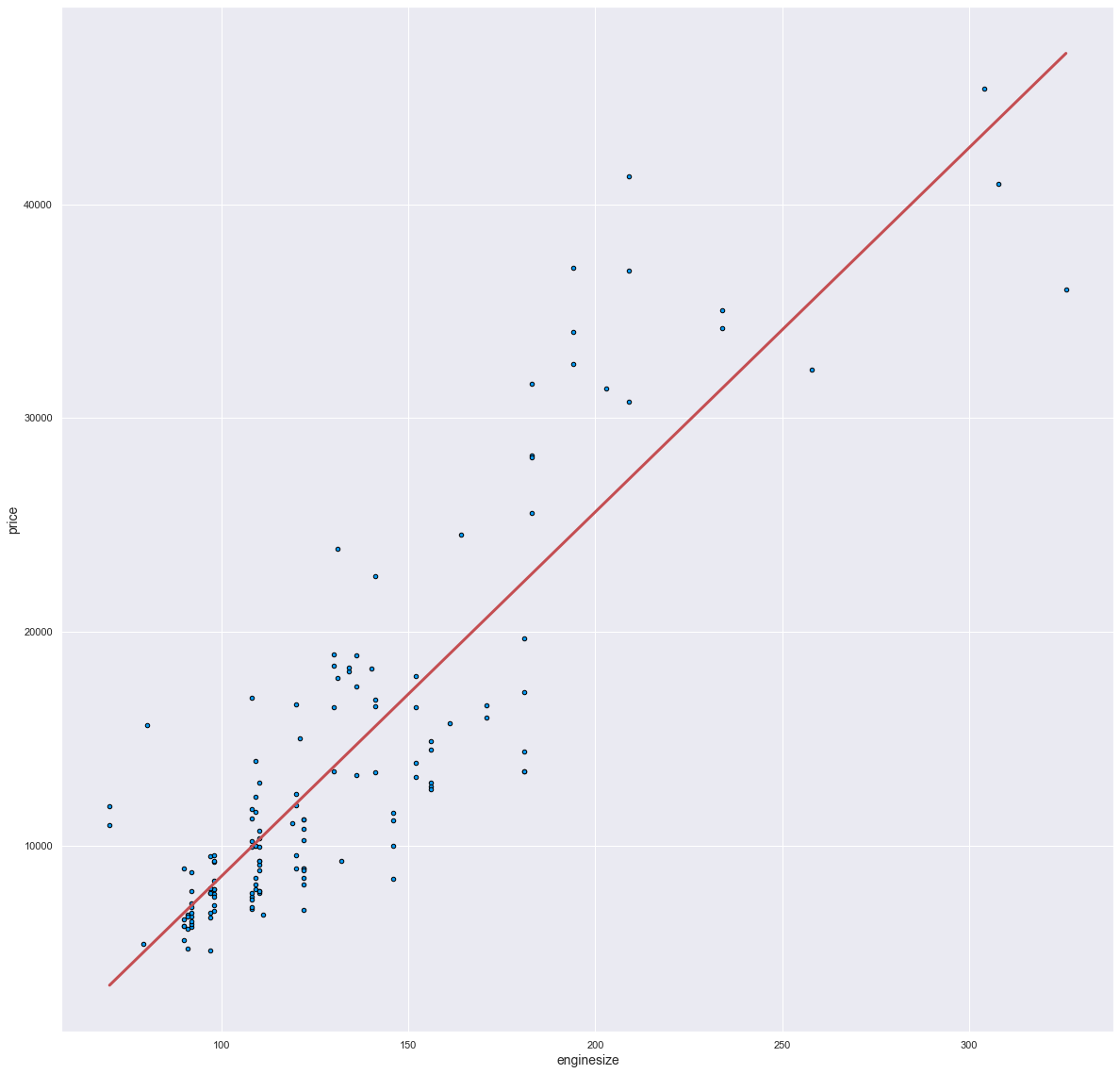


圖5 迴歸模型

如上圖5，此為此次enginesize與price的迴歸模型。

1. **模型效能評估（程式碼Part5. 模型效能評估）：**

此次模型效能評估將以MSE、MAE、作為指標。





圖6 MSE、MAE、

如上圖6，因為MSE過大，因此再以MAE進行評估，並由此可知我們預測的效果可能會很差。再根據的training數據只有0.768且testing數據只有0.730，可知此次迴歸模型確實不夠好。

殘差分析使用Shapiro-Wilk 檢測常態性，因為此次樣本數少於2000，並繪製QQ圖。



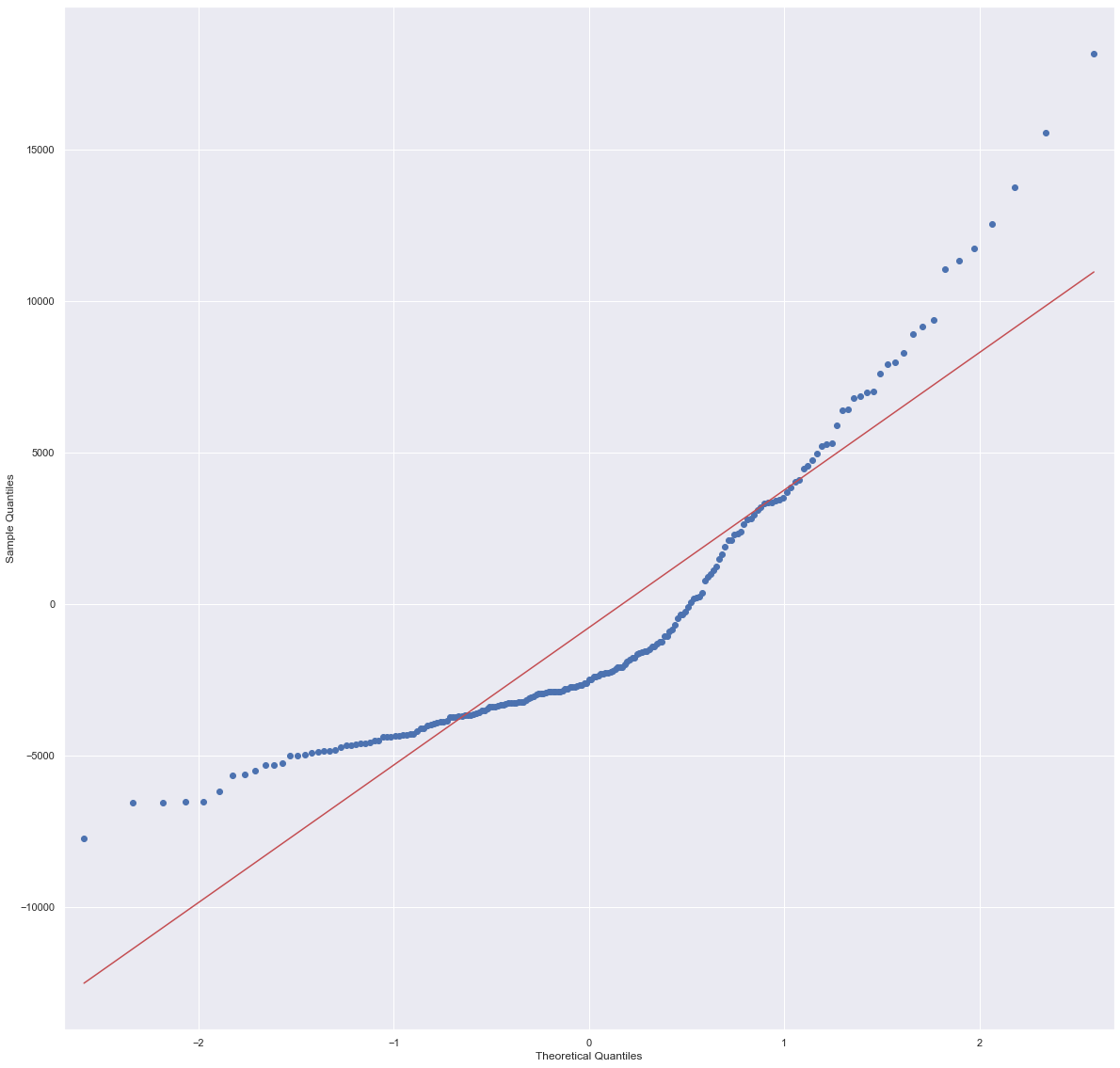


圖7 Shapiro-Wilk、QQ圖

如上圖7，SW檢測出來結果並非常態分佈，QQ圖繪製出來的結果也顯示此次迴歸分析並非常態分佈。

獨立性以Durbin-Watson檢測。



圖8 Durbin-Watson

如上圖8，Durbin-Watson檢測出來結果誤差項並非獨立

變異數同質性以局部加權散點平滑法繪製殘差圖。

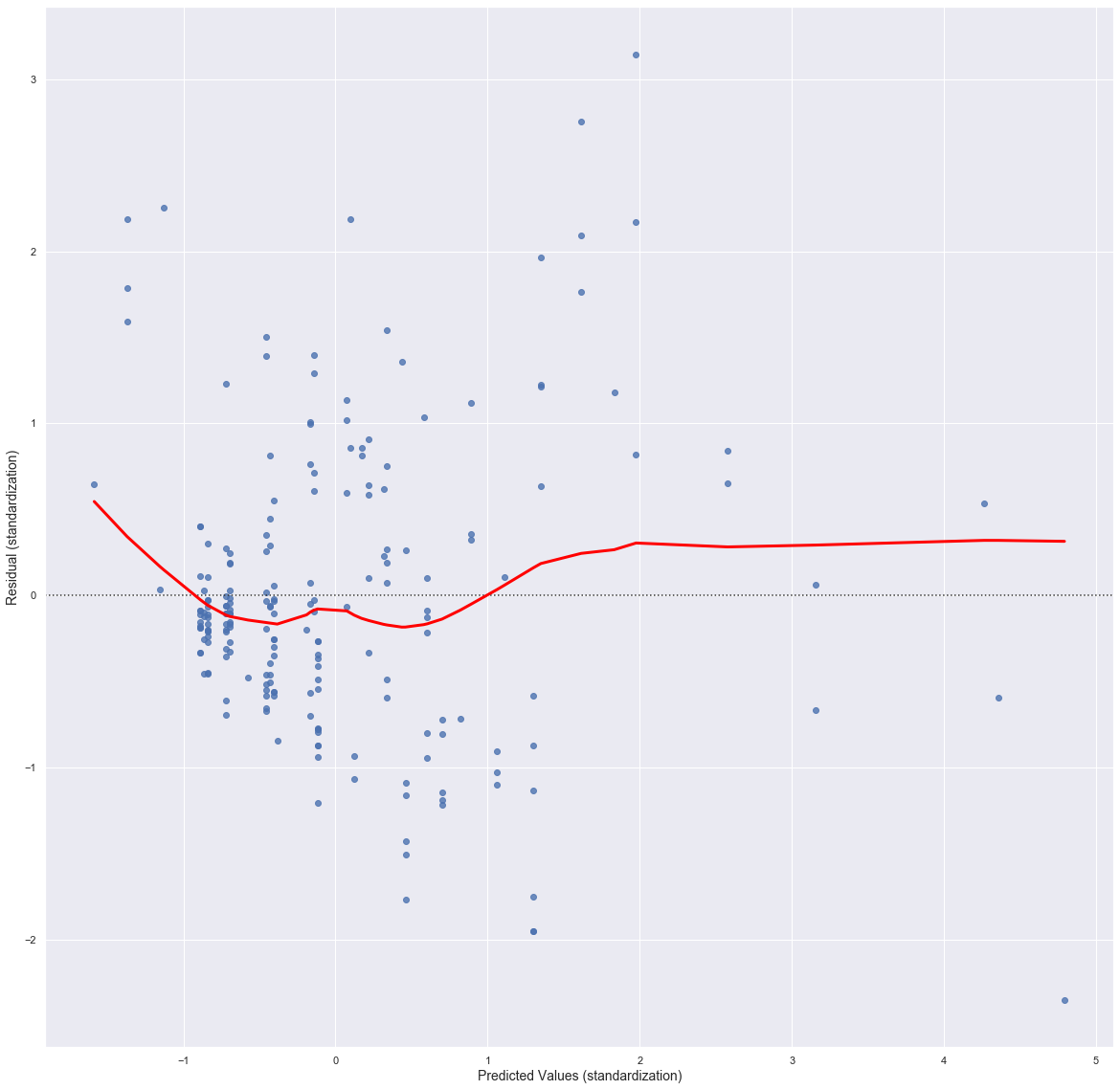


圖9 殘差圖

如上圖9，可由此殘差圖得知，殘差變異數並不符合同質性，因為理想上會希望紅線會靠近0的水平線。

1. **預測結果分析（程式碼Part6. 預測結果分析）：**

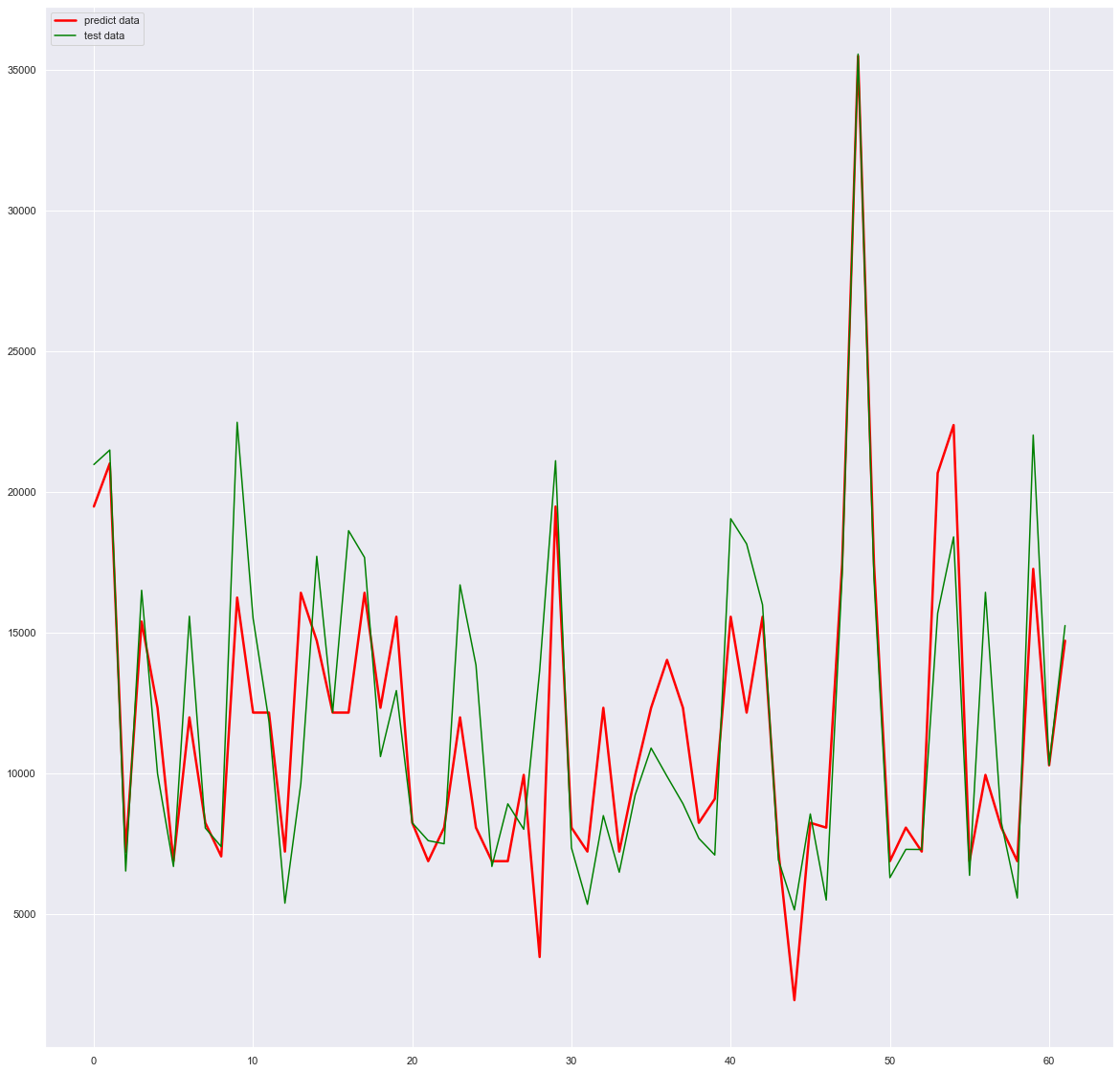


圖10 預測結果分析

如上圖10，經過預測之後，可得知紅線(預測)與綠線(測試集)有極大的落差，可顯現此次的迴歸模型不太理想。

1. **採用多元線性迴歸（程式碼Part7. 採用多元線性迴歸）：**

因為只採用一個自變數的結果不太理想，因此我打算再建立一個模型，並以與Price有高度正相關的carwidth、horsepower、curbweight、enginesize為自變數建立此多元線性迴歸模型。資料分割同樣以7：3的比例切割。

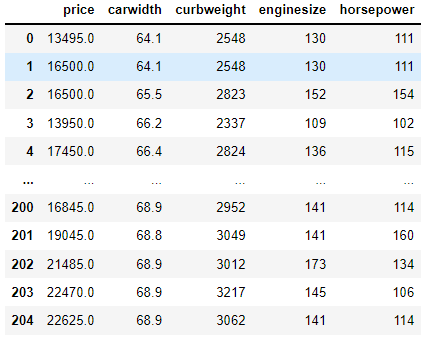


圖11 新資料集

如上圖11，為了方便進行迴歸模型製作，因此將此5個變數另外建新的資料集。

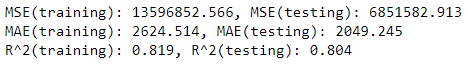


圖12 MSE、MAE、（程式碼30）

如上圖12，MSE與MAE的數值雖然有變小，但仍然很大，由此可知我們預測的效果可能會很差。的testing數值雖然有比上次的0.73高，但仍然未達0.9，因此此次迴歸模型可能還是不夠好。



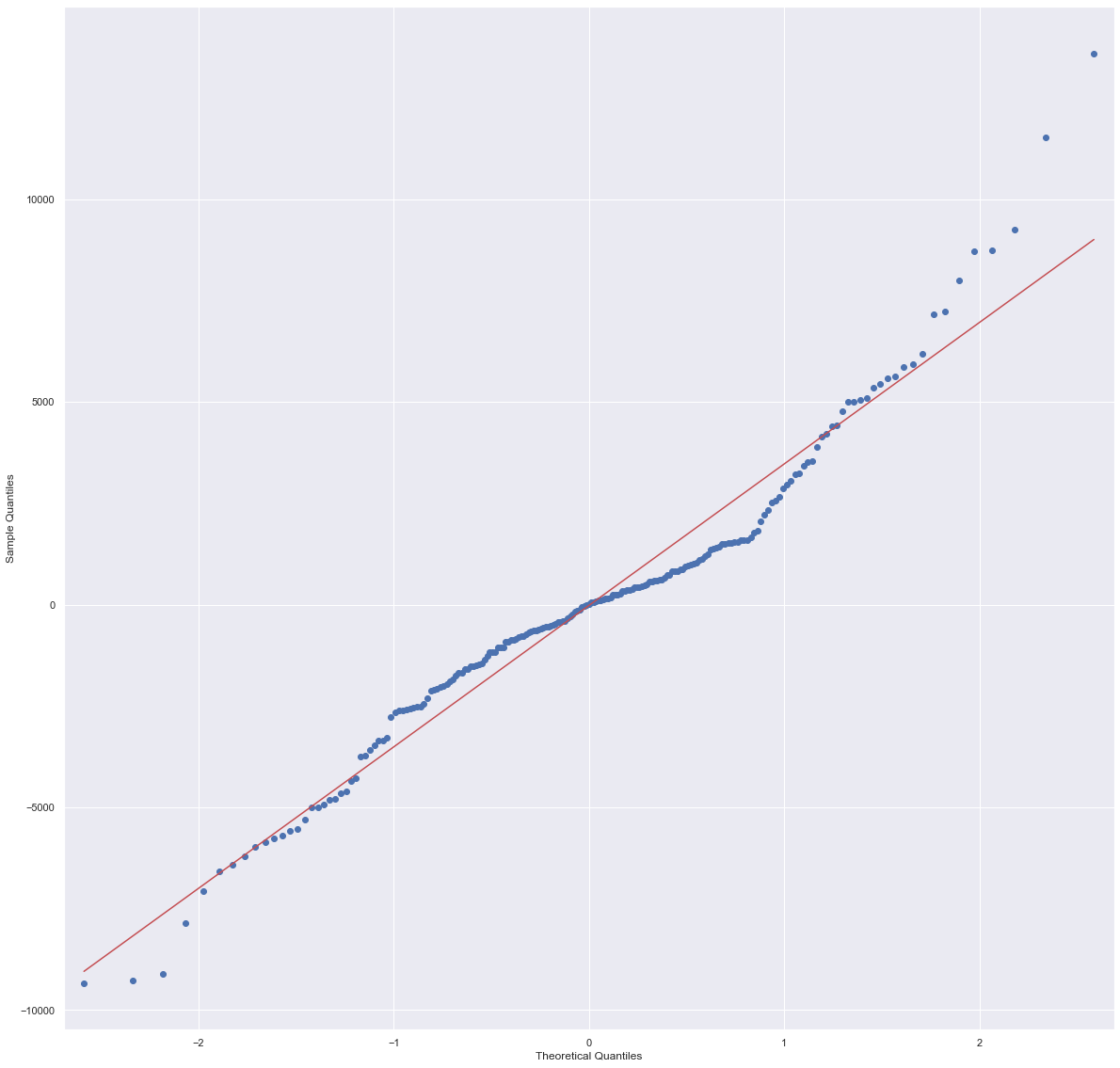


圖13 Shapiro-Wilk、QQ圖（程式碼31）

如上圖13，SW檢測出來結果並非常態分佈，QQ圖繪製出來的結果也顯示此次迴歸分析並非常態分佈，但是QQ圖看起來比起上次更接近常態分佈一點。



圖14 Durbin-Watson（程式碼32）

如上圖14，Durbin-Watson檢測出來結果誤差項並非獨立

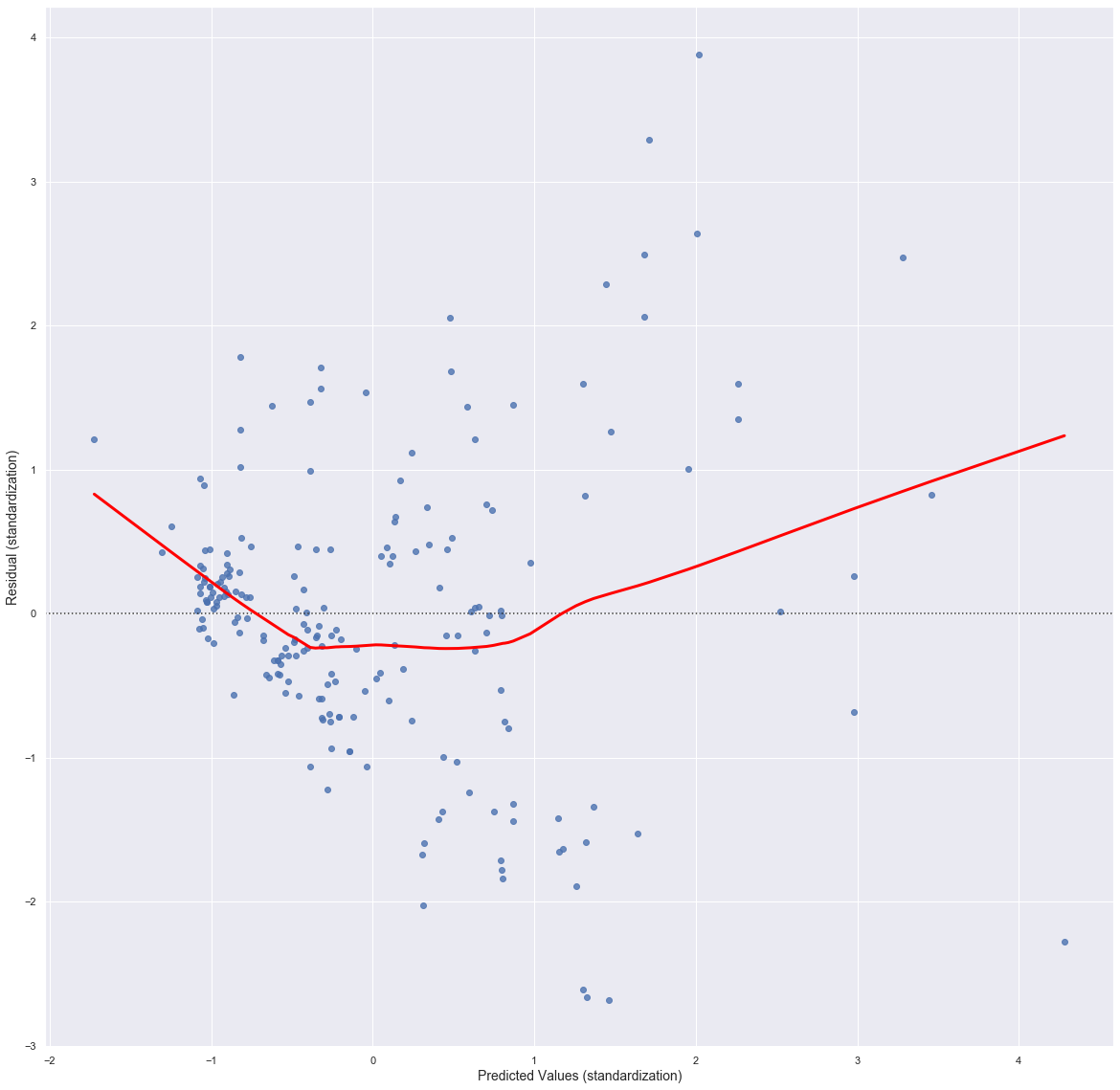


圖15 多元線性迴歸殘差圖（程式碼33）

如上圖15，可由此殘差圖得知，殘差變異數仍不符合同質性，且比起上次的模型，殘差圖有明顯偏離水平線0的位置。

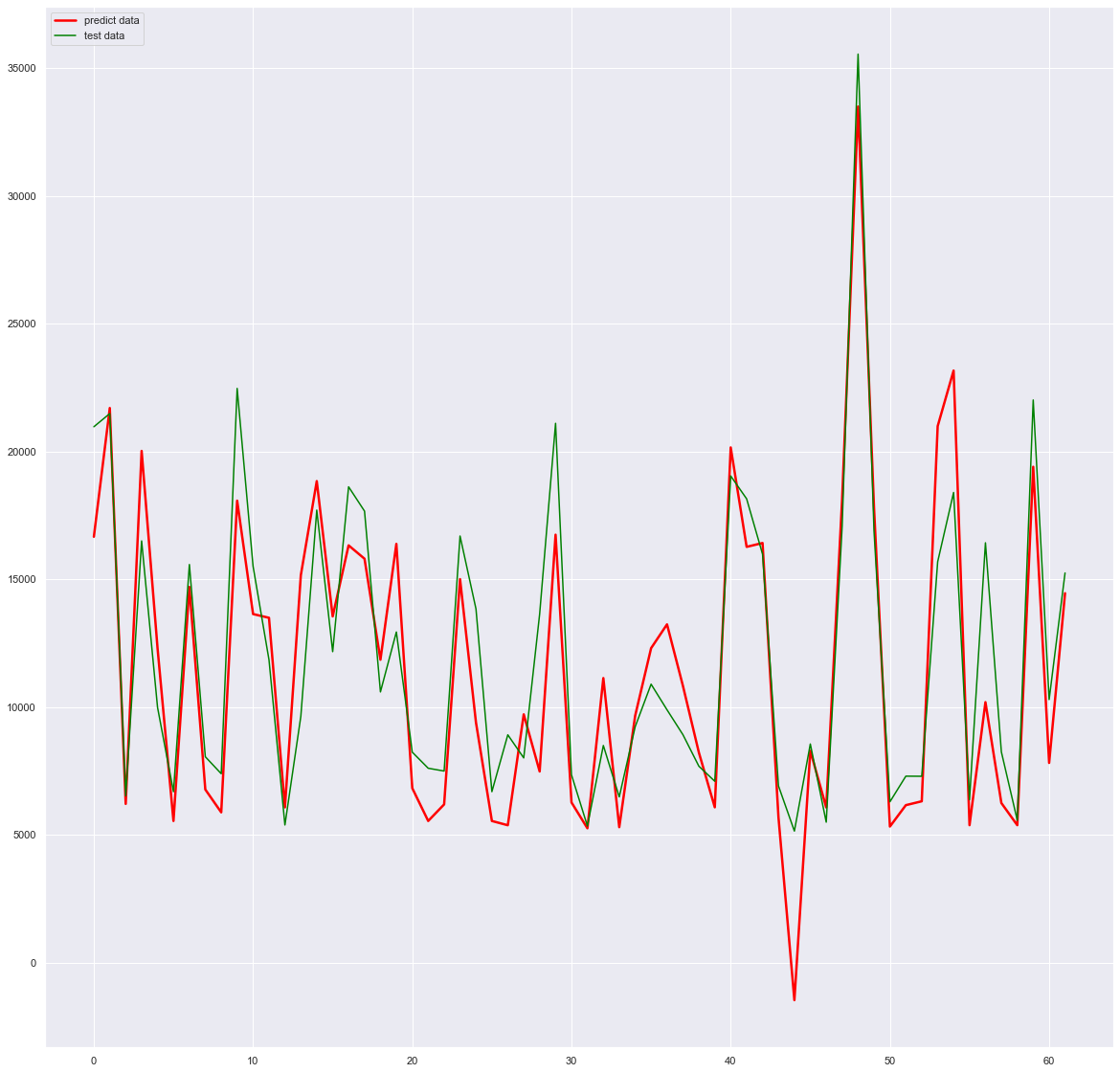


圖16 多元線性迴歸預測圖（程式碼34）

如上圖16，經過預測之後，可得知紅線(預測)與綠線(測試集)雖然只有部分明顯有落差，但是可發現有個點預測的Price小於0，可顯現此次的迴歸模型仍不盡理想。

1. **採用隨機森林迴歸（程式碼Part8. 採用隨機森林迴歸）：**

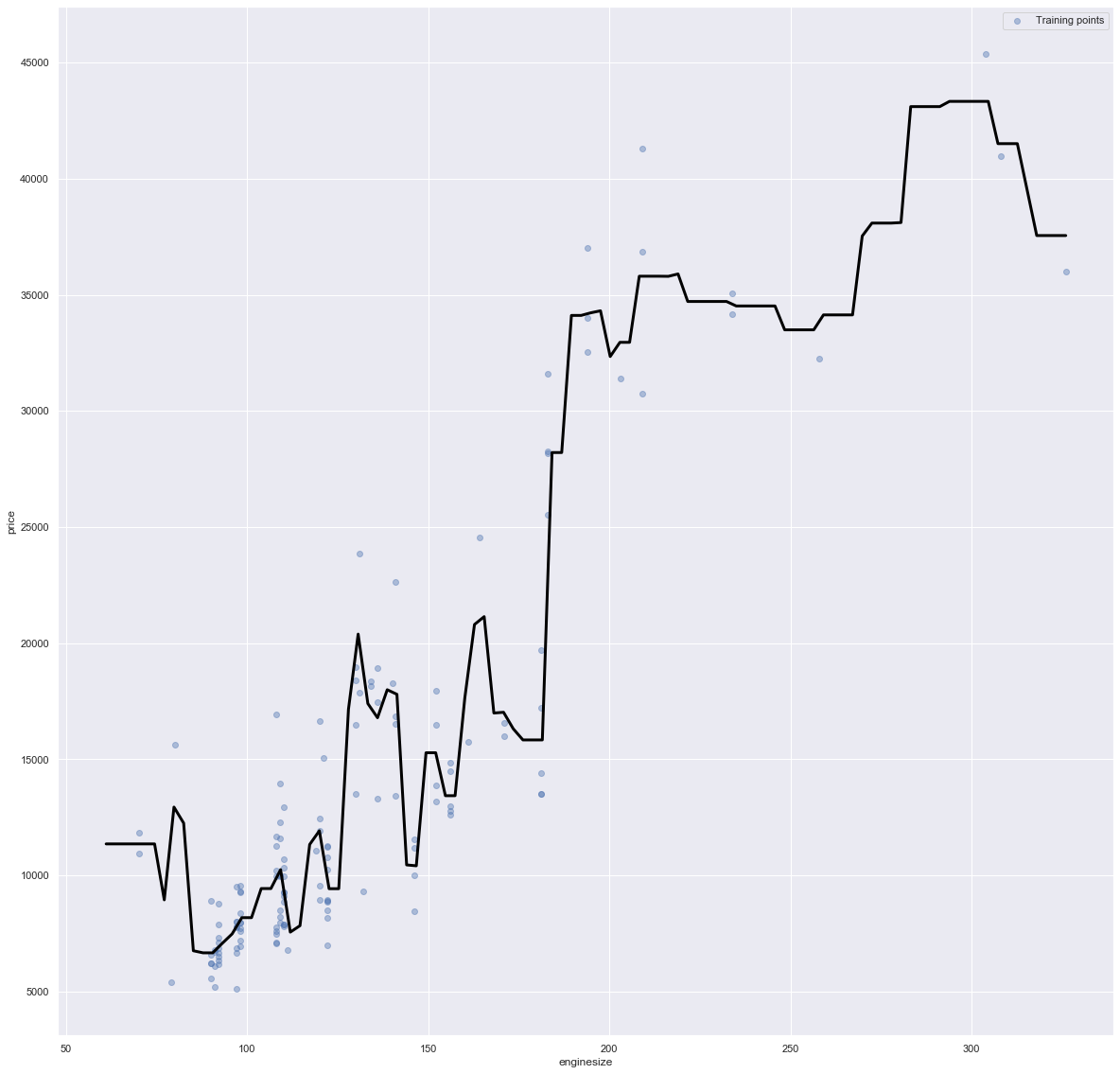
因為多元線性迴歸的結果還是有些許不理想，因此我想再採用隨機森林迴歸模型。因為與Price高度正相關最大的為enginesize，因此本次以enginesize為自變數，Price為應變數建立隨機森林迴歸模型，且資料分割以7：3的比例切割。

圖17 隨機森林迴歸模型（程式碼35）

如上圖17，此為隨機森林迴歸模型建立後，產生的迴歸圖。

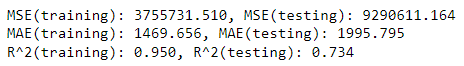


圖18 MSE、MAE、（程式碼36）

如上圖18，MSE與MAE的數值雖然有變更小，但仍然超過1000，由此可知我們預測的效果可能還是會有點差。至於雖然在training方面有0.950，但是在testing只有0.734，因此還是不太理想。



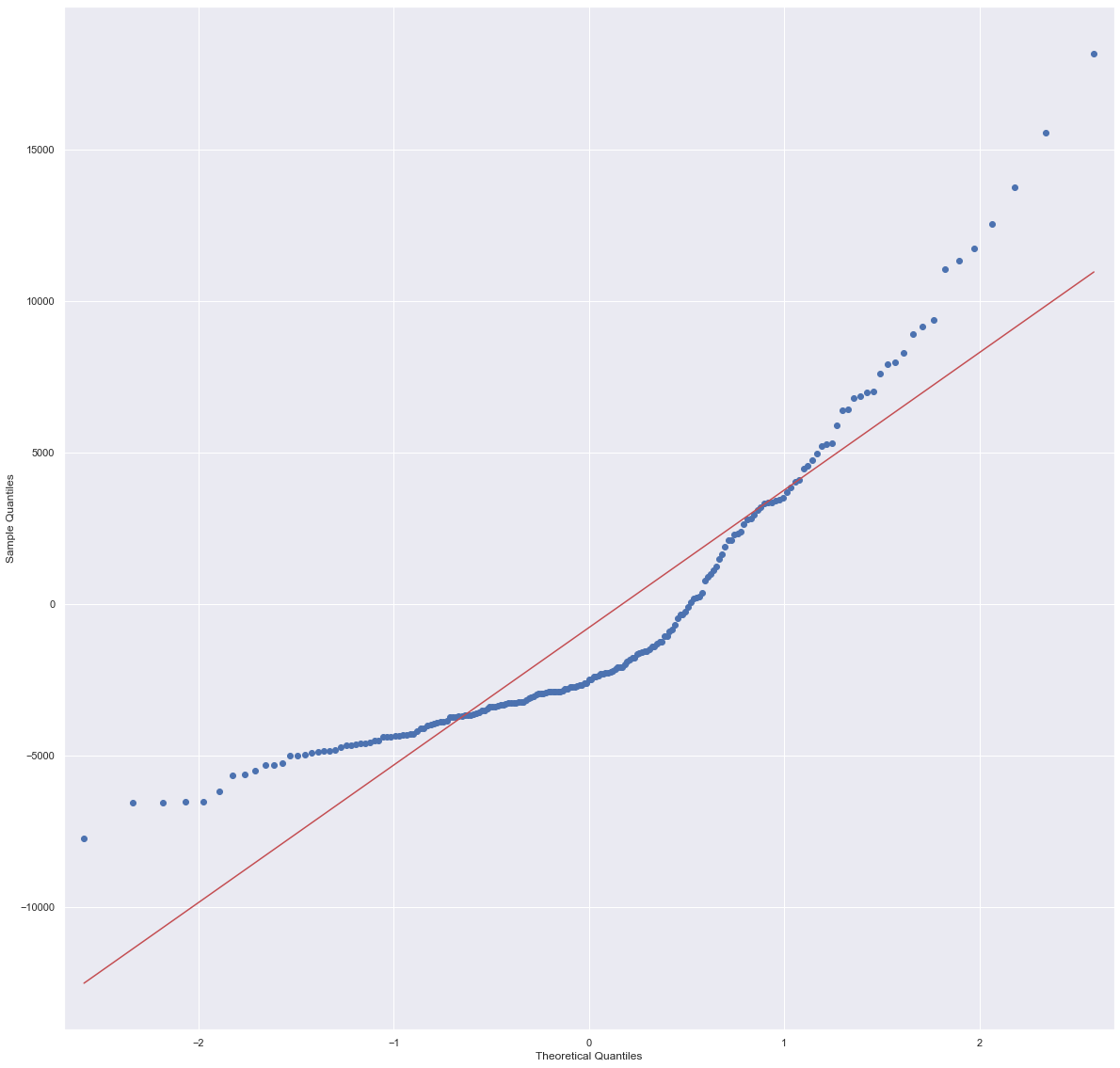


圖19 Shapiro-Wilk、QQ圖（程式碼38）

如上圖19，SW檢測出來結果並非常態分佈，QQ圖繪製出來的結果與普通的線性迴歸看起來很像，因此也顯示此次非常態分佈。



圖20 Durbin-Watson（程式碼39）

如上圖20，Durbin-Watson檢測出來結果誤差項並非獨立

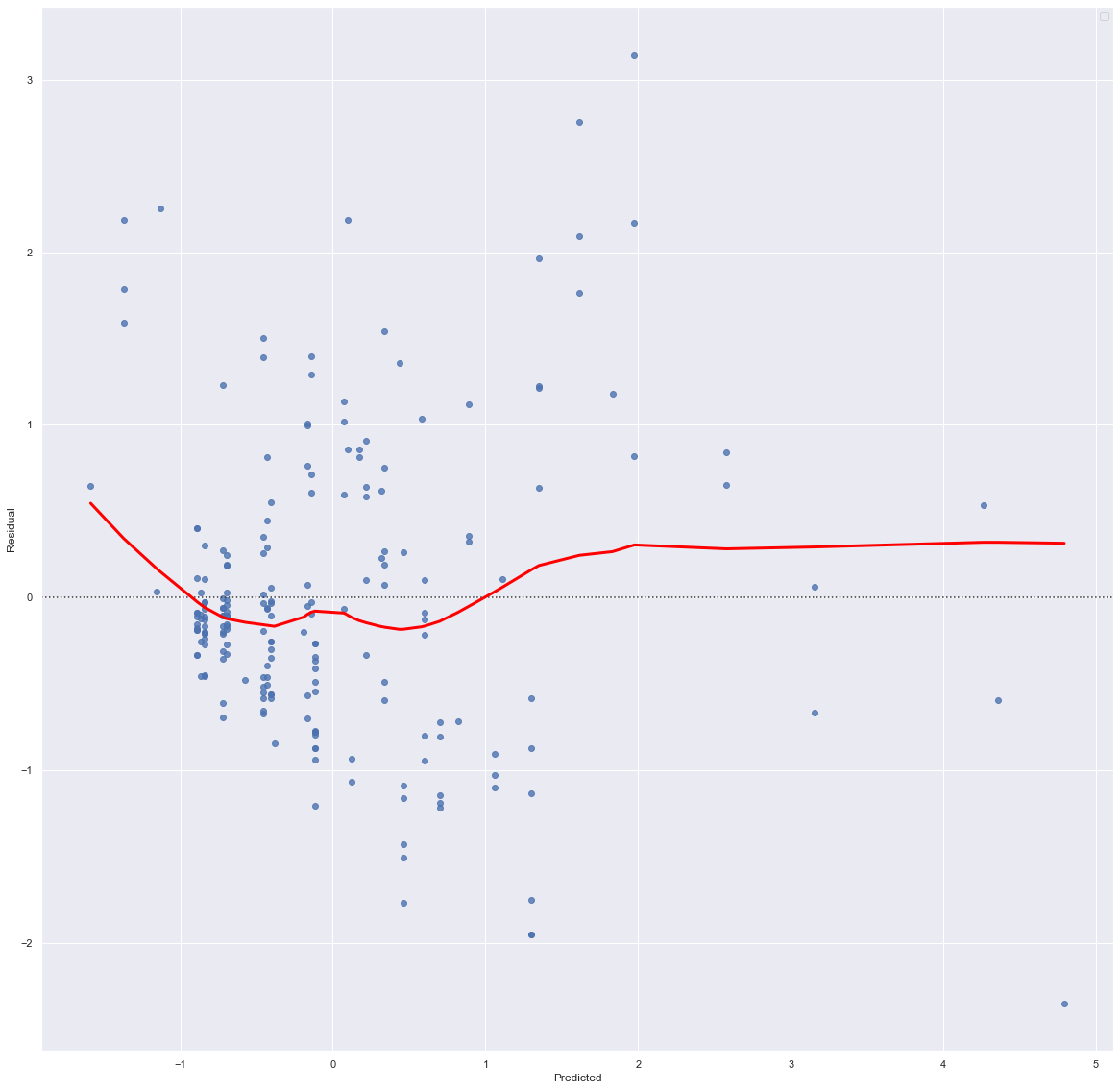


圖21 隨機森林迴歸殘差圖（程式碼40）

如上圖21，此次的殘差圖與普通的線性迴歸看起來也很像，殘差變異數並不符合同質性。

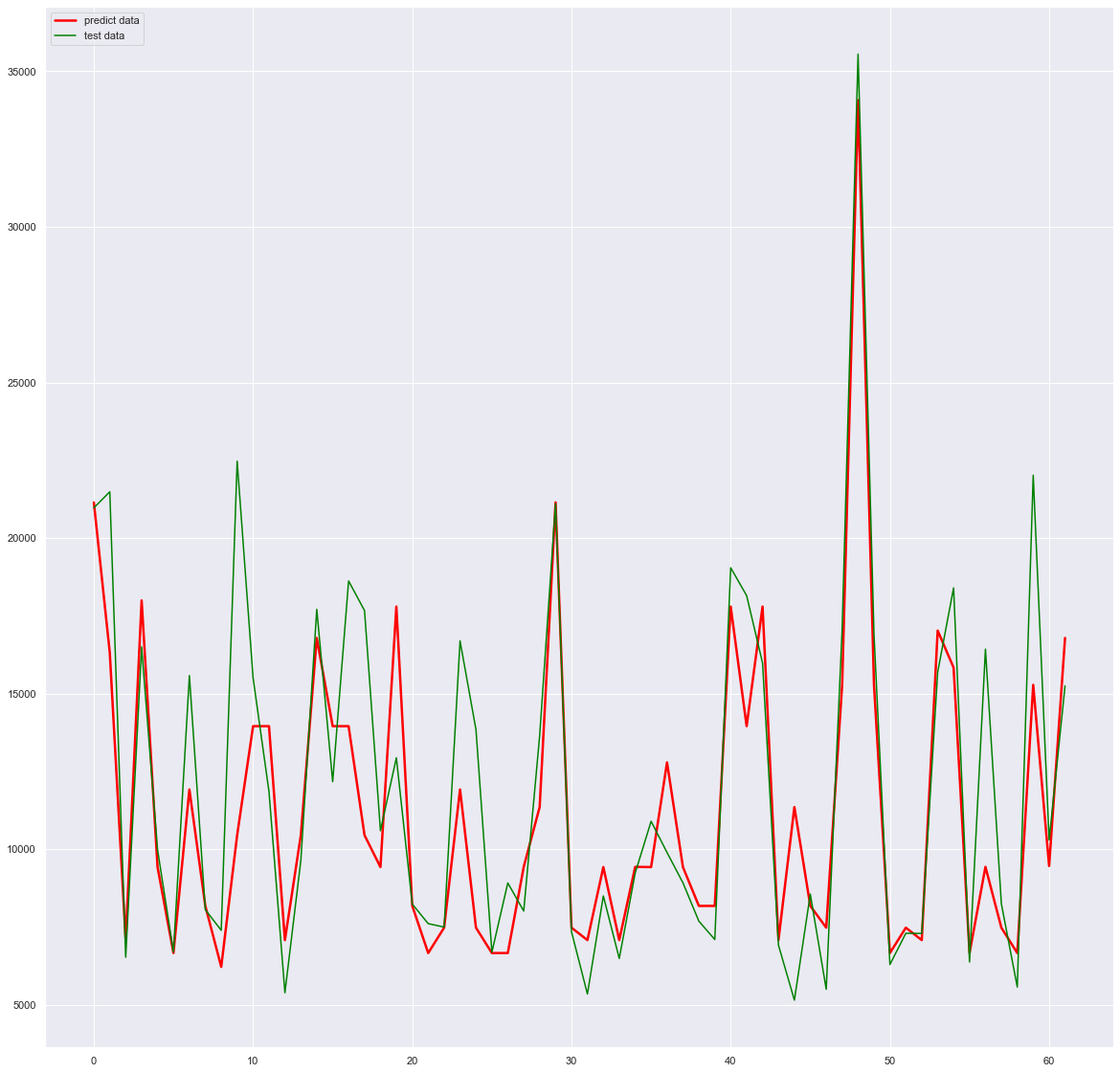


圖22 隨機森林迴歸預測圖（程式碼41）

如上圖22，經過預測之後，可得知紅線(預測)與綠線(測試集)還是有明顯的落差，不過總體來看有比「普通線性迴歸」與「多元線性迴歸」明顯進步。

1. **綜合比較三種模型結果分析：**

|  |  |  |  |
| --- | --- | --- | --- |
| 模型 | 簡單線性迴歸 | 多元線性迴歸 | 隨機森林線性迴歸 |
|  | 0.768 | 0.819 | 0.950 |
|  | 0.730 | 0.804 | 0.734 |

表1 三種模型 比較分析

如上表1，可得知雖然隨機森林在方面有顯著的提升，但是在方面並沒有提升，反而比多元線性迴歸低，才導致此次結果不太理想。

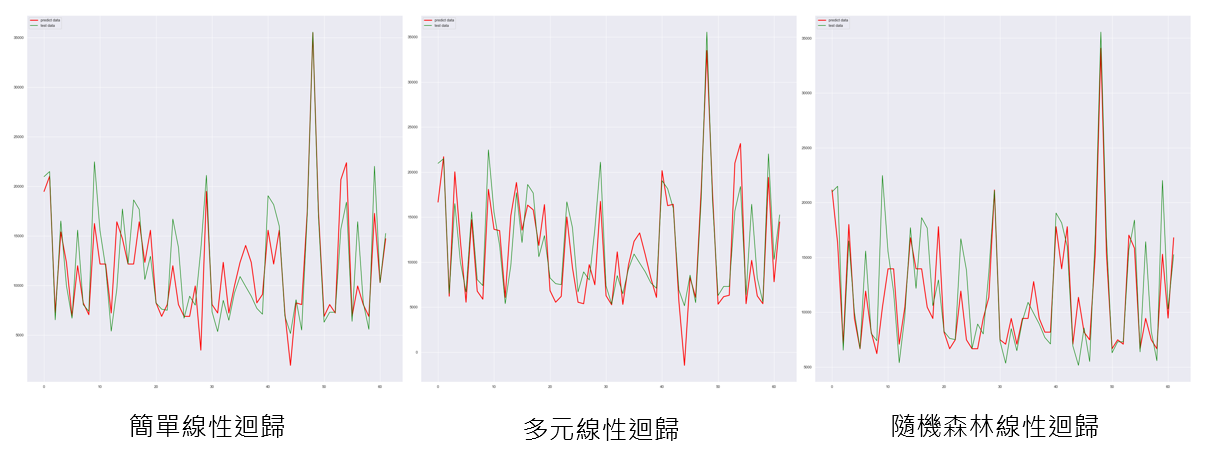


圖23 三種模型預測比較圖

如上圖23，可得知隨機森林線性迴歸的預測結果雖然沒有很好，但是相對於簡單線性迴歸與多元線性迴歸相比仍然有不錯的結果。

對於此次的結果，確實是不太理想，可能是因為我並沒有移除離群值，但是如之前所分析，資料集只有205筆，對於機器學習的資料集來說確實是過少，因此若再刪除離群值可能會讓結果變得更不理想。