自動化網路爬蟲及二手車估價模型

Get in car right now! Here we go!

2020/06-2020/09

專案架構

1

python Linux

Auto Web Crawler
(Selenium、Crontab)

MOBILE01
iCars 車主
8891汽車
abc好車網
HOT大聯盟
u-car.com.tw
yahoo! 中古車

Data Modeling / DataBase
(Data Warebase)

mongoDB
hadoop HDFS
MySQL

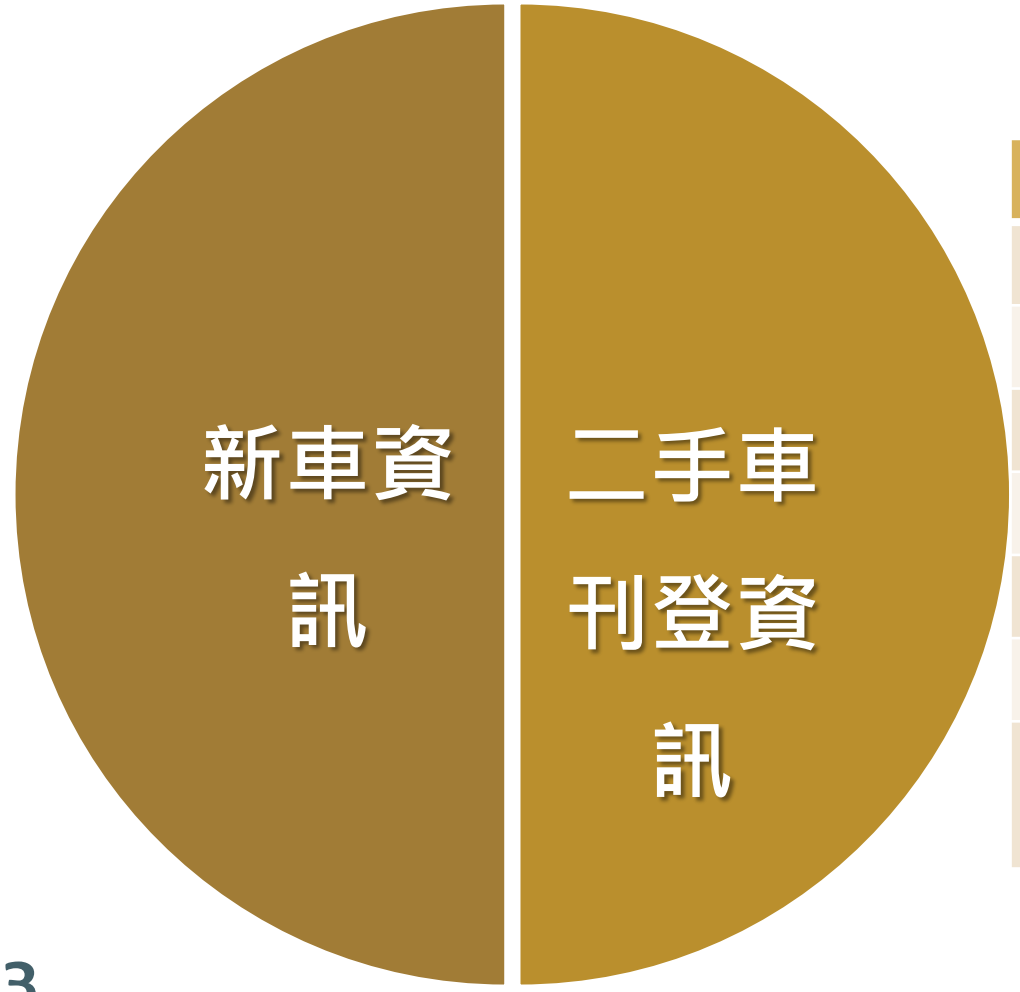
2

ETL (Data Processing)
python R

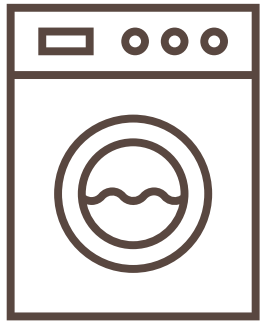
3

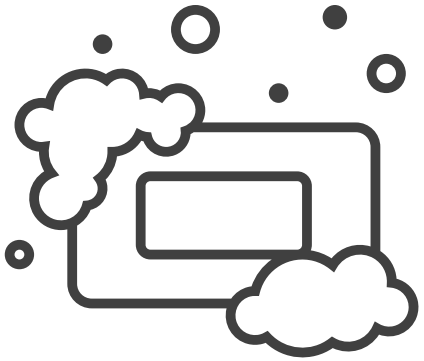
ML
(Data Analysis)
python R

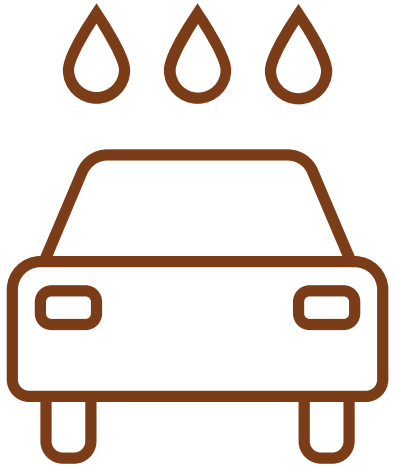
資料來源



資料來源網站名稱	車款資料筆數	圖片張數
中古車王-全國最大二手車資料庫	310,000	約 1,700,000
8891汽車交易網(**新車**、二手車)	39,400	約 520,000
ABC好車網	5,000	約 75,000
Yahoo奇摩中古車	6,500	約 75,000
HOT大聯盟中古車交易網	5,000	約 40,000
U-car(**新車**、二手車)	4,200	約 50,000
小計	**370,100 (約37萬)**	**2,460,000 (約246萬)**







二手車資料清洗

資料清洗

Extract-Transform-Load

挑選欄位

將原始資料從MongoDB取出

Extract

mongoDB

Transform

python

R

Load

MySQL

52

383,869

type	brand	gas	sys	year	price	cc	power	l_chair
A3	Audi	gas	semiauto	22	54	600	2	0
A3	Audi	gas	auto	21	17	1600	2	0
A3	Audi	gas	auto	21	17	1600	2	0
A3	Audi	gas	auto	21	17	1600	2	0
A3	Audi	gas	auto	21	9	1600	2	0
A3	Audi	gas	auto	21	20	1600	2	0
A3	Audi	gas	auto	21	23	1600	2	0
A3	Audi	gas	auto	22	12	1600	2	0
A3	Audi	gas	auto	21	17	1600	2	0
A3	Audi	gas	auto	21	12	1600	2	0
A3	Audi	gas	auto	21	17	1600	2	0
A3	Audi	gas	auto	21	17	1600	2	0
A3	Audi	gas	auto	21	17	1600	2	0

資料清洗

Extract-Transform-Load

Extract

mongoDB

Transform

python

R

Load

MySQL

挑選欄位

↓

清洗資料

↓

處理遺漏值

1. 針對各欄位遺漏值進行交叉比對，同一車型以該值最大佔比進行補值

2. 剔除空值過半的欄位

里程數 85%資料為遺漏值

```
> sum(is.na(sub$miles))/dim(sub)[1]
[1] 0.8576989
```

3. 剩餘空值，該欄位以最大佔比值進行補值

傳動數 2輪傳動佔資料比79%
遺漏值以2輪傳動進行補值

```
> sum(power==2)/dim(sub)[1]
[1] 0.7937968
```

資料清洗

Extract-Transform-Load

Extract

mongoDB

Transform

python

R

Load

MySQL

挑選欄位

清洗資料

處理遺漏值

去除雜訊

資料經ETL後剩19%，有效欄位24，總筆數160,137

```
> dim(data)[1]*dim(data)[2]/(dim(raw)[1]*dim(raw)[2])
[1] 0.192538
> dim(data)
[1] 160137     24
```

原始資料經ETL後存入MySQL

24

160137

type	brand	gas	sys	year	price	cc	power	l_chair
A3	Audi	gas	semiauto	22	54	600	2	0
A3	Audi	gas	auto	21	17	1600	2	0
A3	Audi	gas	auto	21	17	1600	2	0
A3	Audi	gas	auto	21	17	1600	2	0
A3	Audi	gas	auto	21	9	1600	2	0
A3	Audi	gas	auto	21	20	1600	2	0
A3	Audi	gas	auto	21	23	1600	2	0
A3	Audi	gas	auto	22	12	1600	2	0
A3	Audi	gas	auto	21	17	1600	2	0
A3	Audi	gas	auto	21	12	1600	2	0
A3	Audi	gas	auto	21	17	1600	2	0
A3	Audi	gas	auto	21	17	1600	2	0
A3	Audi	gas	auto	21	17	1600	2	0

二手車
估價預測

售車估價

欲售二手車 售價預測

建立預測模型 1

挑選合適變數

建立預測模型 2

因價格為連續性數值，選以監督式學習(CART)決策樹、(Linear Regression)多元複回歸模型分析。

	Linear Regression
Adjusted R square	76.69%
精準度評估MAPE	41.03%

year	-5.012e+00	2.623e-02	-191.098	< 2e-16	***
cc	2.183e-02	1.933e-04	112.885	< 2e-16	***
power	5.500e-01	2.286e-01	2.406	0.0161	*
l_chair	-2.645e+00	3.039e-01	-8.703	< 2e-16	***
auto_chair	2.628e+01	3.881e-01	67.727	< 2e-16	***
back_screen	-7.831e+00	4.436e-01	-17.653	< 2e-16	***
back_radar	-1.180e+00	4.624e-01	-2.552	0.0107	*
ABS	-1.134e+01	4.641e-01	-24.429	< 2e-16	***
window	8.765e+00	3.343e-01	26.219	< 2e-16	***
hid	-2.647e+00	3.403e-01	-7.778	7.41e-15	***
safe_bag	4.030e+00	3.094e-01	13.025	< 2e-16	***
gps	-6.099e+00	3.984e-01	-15.308	< 2e-16	***
keyless	-1.810e+00	4.146e-01	-4.366	1.27e-05	***
led	-2.855e+00	4.141e-01	-6.895	5.42e-12	***
tcs	1.692e-01	3.002e-01	0.563	0.5731	

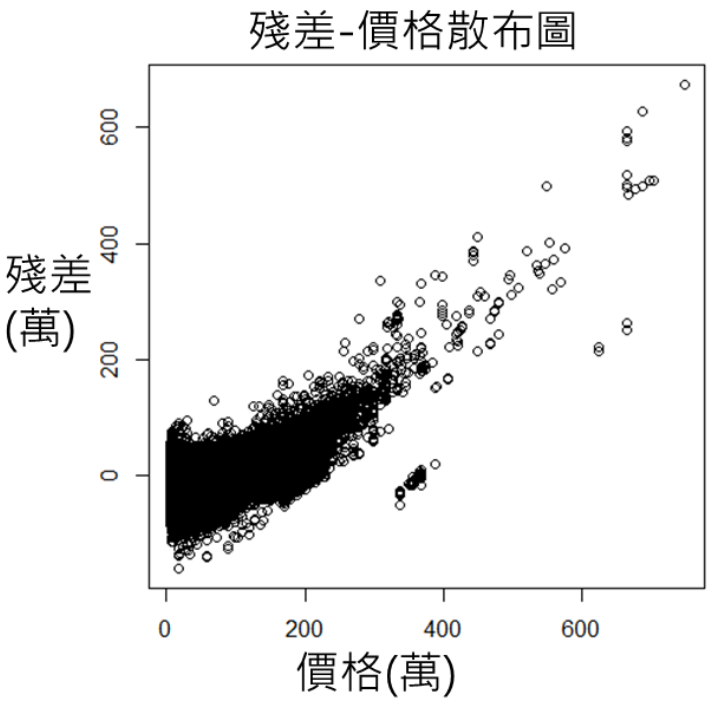
售車估價

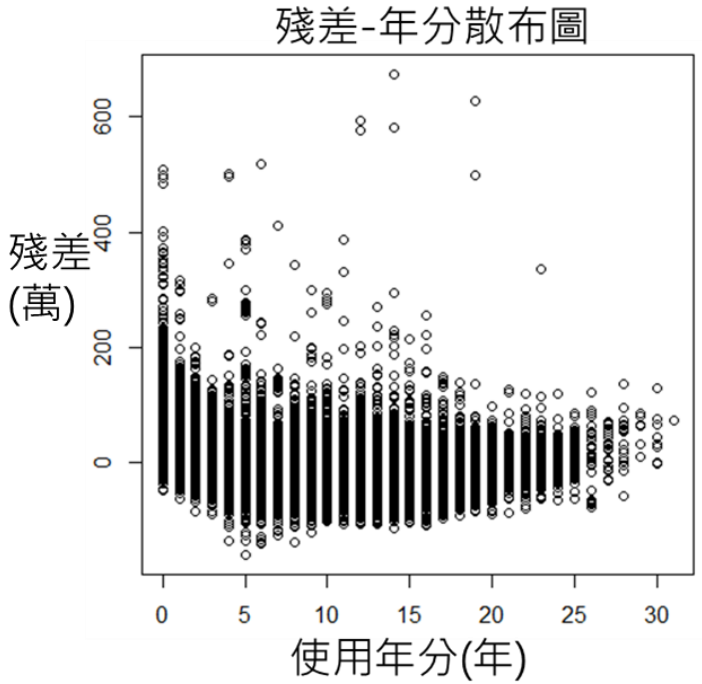
殘差分布

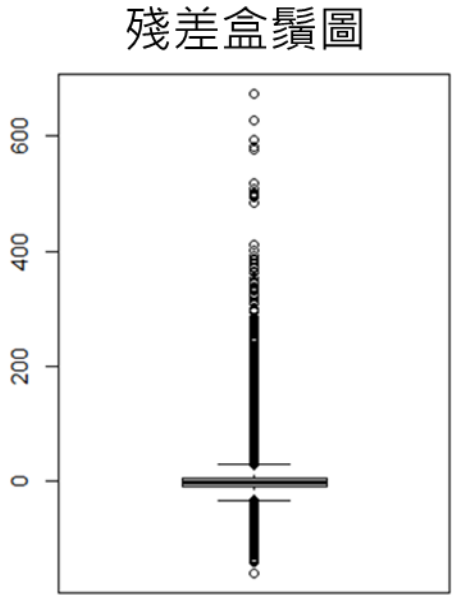
觀察殘差

剔除離群值

新資料集








```
> quantile(res)
          0%          25%          50%          75%         100%
-158.5543312   -8.4432158   -0.7559996    7.0658341  673.2205511
> sd(res)
[1] 21.6326
> IQR(res)
[1] 15.50905
```

售車估價

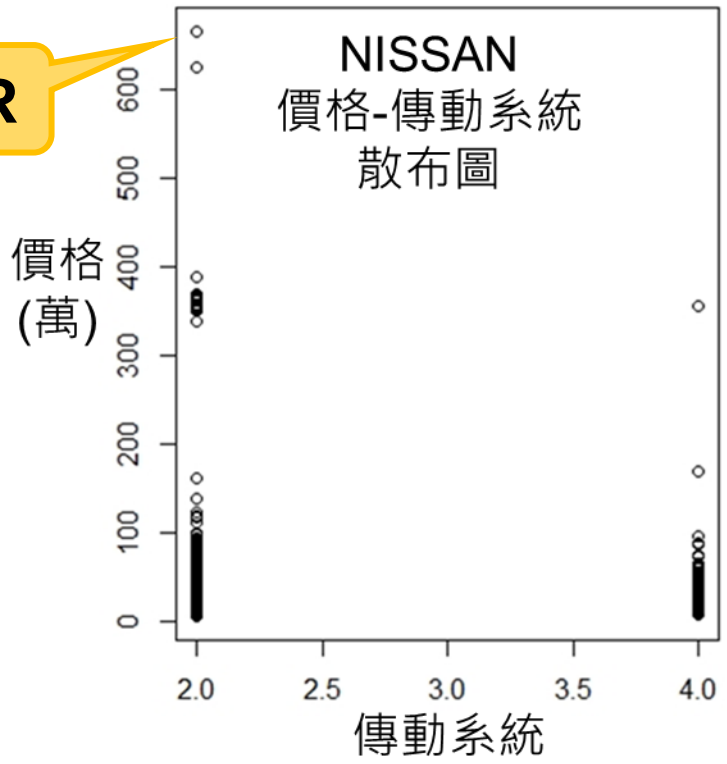
欲售二手車 售價預測

建立預測模型 3

剔除不適變數

```
#使用者二手車市價估算 複回歸方程式
price_predict=168.106775+11.54561071*car["auto_chair"]-1.072887273*car["safe_bag"]\
      -4.56611175*car["power"]-0.001975656*car["cc"]-3.371553644*car["year"]\
       -2.52933161*car["semiauto"]-2.705297375*car["hand"]+16.41229107*car["hybrid"]\
       +9.246868102*car["gas"]+1*car["coeb"]+1*car["coet"]
#估算合理價上下限，殘差值的標準差6.684836，單位萬
estimate_price=round(price_predict,2)
upboard=round(estimate_price+6.68,1)
downboard=round(estimate_price-6.68,1)
```

GTR



售車估價

欲售二手車 售價預測

建立預測模型 3

剔除不適變數

建立預測模型 4

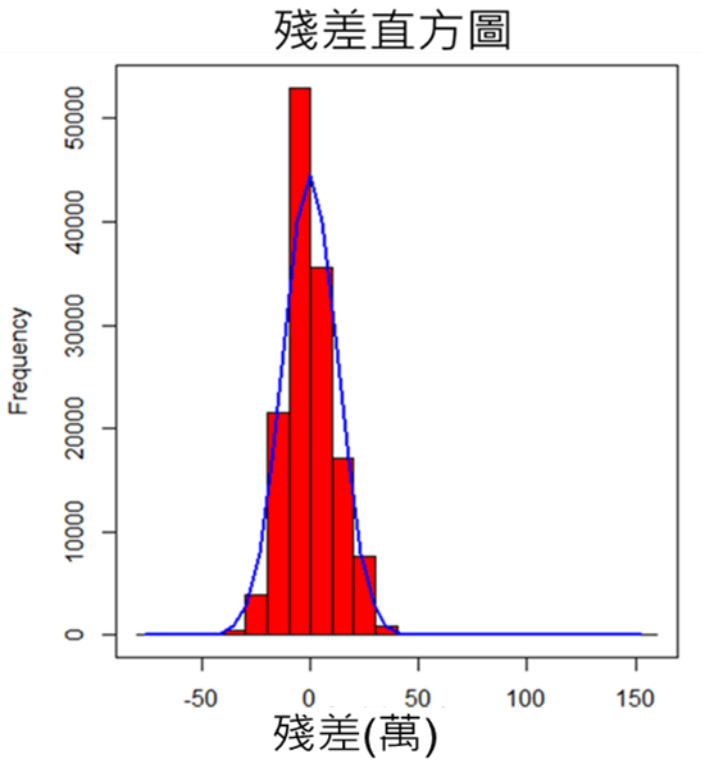
觀察殘差

```
#使用者二手車市價估算 複回歸方程式
price_predict=154.1392221+11.80682461*auto_chair-6.41950482*l_chair\
      +4.571555743*safe_bag+1.220113211*window-0.002609202*cc\
    -3.37803454*year-3.76528218*semiauto-4.720416481*hand\
      +16.11014399*hybrid+8.901650607*gas+1*coe_dic[brand]+1*coe_dic[type]
#估算合理價上下限，殘差值的標準差6.684836，單位萬
estimate_price=round(price_predict,2)
upboard=round(estimate_price+7.05,1)
downboard=round(estimate_price-7.05,1)
print("合理價區間:",downboard,"~",upboard)
```

```
> IQR(absres)
[1] 8.411362
> s2=sd(absres)
> s2
[1] 7.050192
```

絕對殘差 四分位距

絕對殘差 標準差

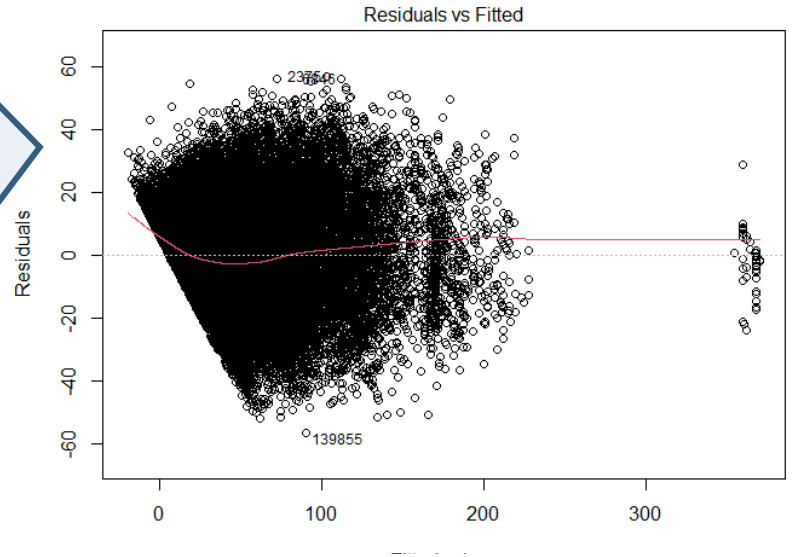


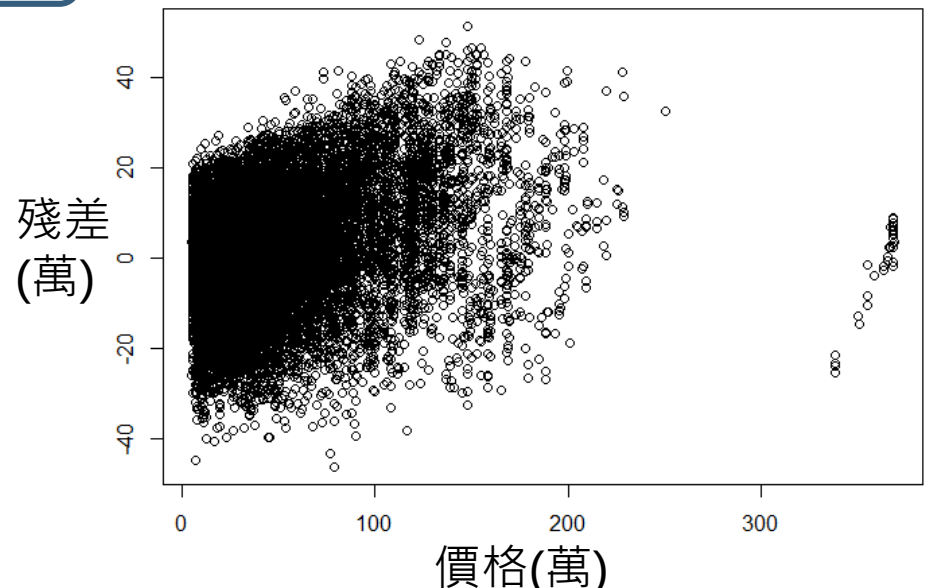
售車估價

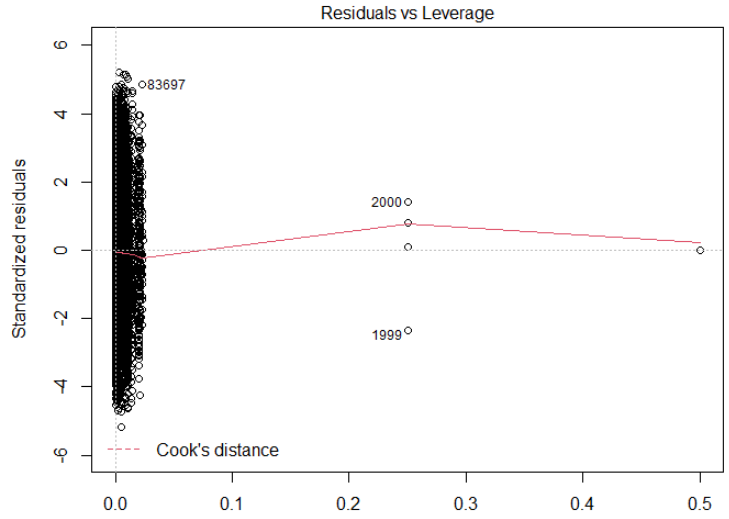
殘差分布

1. 建立預測模型 3
2. 剔除不適變數
3. 建立預測模型 4
4. 觀察殘差

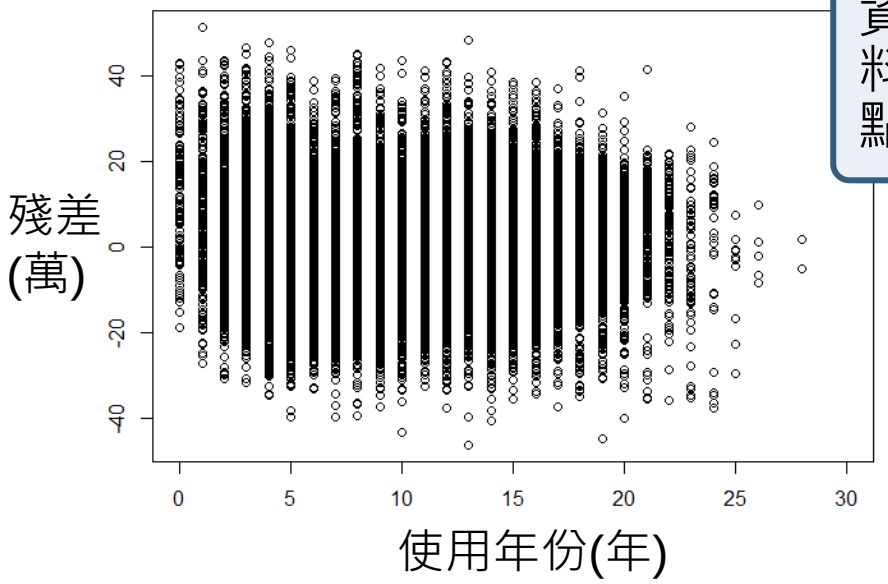
殘差變異數符合均一性







無對模型影響力過高的資料點



售車估價

欲售二手車 售價預測

建立預測模型 3

↓

剔除不適變數

↓

建立預測模型 4

↓

觀察殘差

	Linear Regression多元複回歸	CART 決策樹
精準度評估MAPE	41.03%	49.73%
精準度評估MAPE (剔除離群值)	29.42%	43.09%
精準度評估MAPE (加上區間值)	6.64%	11.01%

```
> test.MAPE
[1] 0.06639495
```

```
> IQR(absres)
[1] 8.411362
> s2=sd(absres)
> s2
[1] 7.050192
```

絕對殘差 四分位距

絕對殘差 標準差

售車估價

售價預測系統比較

```
enter brand
Nissan
enter type
Livina
有電動椅輸入1 沒有輸入0
0
有安全氣囊輸入1 沒有輸入0
1
有天窗輸入1 沒有輸入0
0
輸入排氣CC數
1600
輸入西元出廠年份
2015
手自排車輸入1 否輸入0
0
油電混和車輸入1 否輸入0
0
合理價區間: 29.9 ~ 44.0
```

售車估價

售價預測系統比較

估價網站	估價(萬)	平均值(萬)
資料庫(138筆)	28.0~35.6	30.6
老司機估價模型	**29.9~44.0**	**36.9**
SUM汽車網	29.8~36.8	33.3
CARP汽車鑑價網	33.5	33.5
CARGURU車咕嚕	29.5~31.9	30.7
ABC好車網	88.0~95.0	91.5
AUTOSTAR	27.0~32.0	29.5