**探討耗油量、排氣量與車子的性能配件的關係**

410978004 統計三 郭依璇

目錄

1. 背景、動機及研究目的
2. 資料說明
3. 多變量分析方法
4. 分析結果
5. 綜合結論及建議
6. 參考文獻

一、背景、動機及研究目的

在這個幾乎人人都會開車、環保意識抬頭的年代，年輕人都會在意何種車的耗油量會較少、可以少用點油而達到環保的效果。此外，還在乎排氣量，排越少的廢氣，就能對環境更好。於是此研究要探討，何種車的耗油量較少、排氣量較少。

二、資料說明

| **Var** | **Description** | **Comments** |
| --- | --- | --- |
| mpg  (metric) | Miles/US Gallon | 燃油效率的決定因素 |
| cyl  (nonmetric) | Number of cylinders | 數據包括配備 4、6、8 氣缸數的車輛。 |
| disp  (metric) | Displacement (cu.in.) | 排量測量發動機中的總容積，作為氣缸周長、深度和氣缸總數的一個因素。 |
| hp  (metric) | Gross horsepower | 總馬力衡量的是發動機功率輸出的理論輸出 |
| drat  (metric) | Rear axle ratio | 後橋傳動比表示輪軸每轉一圈驅動軸的轉數。 |
| wt  (metric) | Weight (lb/1000) | 每 1000 磅車輛的總重量（半美噸） |
| qsec  (metric) | 1/4 mile time | 一種性能測量，主要是加速。從靜止狀態行駛 1/4 英里的最快時間（以秒為單位）。 |
| vs  (nonmetric) | V/S | 指示發動機氣缸配置的二進制變量為 V 形 (vs=0) 或直線 (vs=1)。 |
| am  (nonmetric) | Transmission Type | 一個二進制變量，自排車 (am=0) 或手排車(am=1)。 |
| gear  (nonmetric) | Number of forward gears | 變速器中的齒輪數。手動變速箱有 4 個或 5 個前進檔；自動 3 或 4 |
| carb  (nonmetric) | Number of carburetors | 化油器桶的數量。該指標是化油器數量和化油器內桶數的總和。 |

資料共有32筆(obs=32)

mtcars的資料來自於1974 Motor Trend US雜誌，也是R內建有的資料，包含32台車與11個相關的參數

Y : 車子性能

= mpg

= disp

X : Car measures

= qsec

= hp

= drat

= wt

B : background

= vs

= am

= gear

= carb

= cyl

資料來源 : <https://gist.github.com/seankross/a412dfbd88b3db70b74b>

三、多變量分析方法

Step1 : 用canonical看Y與X構面有無相關

Step2 : 將、、做cluster，可看出有低mpg、低disp的X構面的特徵

Step3 : 把cluster當作類別(裡面有cluster=1 cluster=2)，background對cluster進行manova分析，可得知rejH0的background與不同cluster是有差異的，也就是說，此background和mpg、disp有關係，可以做進一步的分析

Step4 : 建立複回歸模型y為、，predictors為和上一步測出來顯著的background

Step5 : 用backward篩選出最終會和mpg disp有顯著相關的模型

Step6 : 看最終模型的各種統計量並進行分析

Step7 : 綜合以上步驟進行總結

四、分析結果

canonical

= mpg

= disp

= qsec

= hp

= drat

= wt

: 0.4381 – 0.6013

: 0.1860 – 0.2003 + 0.1764 - 0.6522

= 0.97

: -1.8326 – 1.7857

: 1.0527 – 1.3515 + 1.0901 – 0.1707

= 0.04

Test for the significance of the canonical relations with

vs

F = 20.29 ; p-value < 0.0001 < 0.05 → reject at , at least the first pair of canonical variables are important.

vs

F = 0.02 ; p-value = 0.9969 > 0.05 → do not reject at , only the first pair of canonical variables are important.

Thus, the first canonical variates are good summary measures.

: mpg、disp表現佳，故命名為車子性能指標

: wt表現佳，故命名為weight指標

RM(Y|X) = (0.8454+0.8902)/2 = 0.8687

RM(X|Y) = (0.1974+0.6653+0.5267+0.8355)/4 = 0.5562

Y構面可被X構面解釋約87%

* 也就是說，X跟Y構面有相關，可做後續分析

cluster

Cluster = 1

| **變數** | **N** | **平均值** | **標準差** | **最小值** | **最大值** |
| --- | --- | --- | --- | --- | --- |
| |  | | --- | | **mpg** | | **disp** | | **hp** | | **drat** | | **wt** | | **qsec** | | |  | | --- | | 16 | | 16 | | 16 | | 16 | | 16 | | 16 | | |  | | --- | | 15.6812500 | | 339.1500000 | | 196.5000000 | | 3.1906250 | | 3.9165000 | | 17.1543750 | | |  | | --- | | 2.9267089 | | 73.9587092 | | 58.8217647 | | 0.3670689 | | 0.7435644 | | 1.5332622 | | | |  | | --- | | 10.4000000 | | 225.0000000 | | 105.0000000 | | 2.7600000 | | 3.1700000 | | 14.5000000 | | |  | | --- | | 21.4000000 | | 472.0000000 | | 335.0000000 | | 4.2200000 | | 5.4240000 | | 20.2200000 | |

Cluster = 2

| **變數** | **N** | **平均值** | **標準差** | **最小值** | **最大值** |
| --- | --- | --- | --- | --- | --- |
| |  | | --- | | **mpg** | | **disp** | | **hp** | | **drat** | | **wt** | | **qsec** | | |  | | --- | | 16 | | 16 | | 16 | | 16 | | 16 | | 16 | | |  | | --- | | 24.5000000 | | 122.2937500 | | 96.8750000 | | 4.0025000 | | 2.5180000 | | 18.5431250 | | |  | | --- | | 5.0026660 | | 34.5671997 | | 31.0115570 | | 0.3233265 | | 0.6185472 | | 1.7943568 | | |  | | --- | | 17.8000000 | | 71.1000000 | | 52.0000000 | | 3.6200000 | | 1.5130000 | | 15.5000000 | | |  | | --- | | 33.9000000 | | 167.6000000 | | 175.0000000 | | 4.9300000 | | 3.4400000 | | 22.9000000 | |

Cluster1與Cluster2的比較

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | mpg | disp | hp | drat | wt | qsec |
| Cluster1 | 低 | 高 | 高 | 低 | 高 | 低 |
| Cluster2 | 高 | 低 | 低 | 高 | 低 | 高 |

* 由此可看出，mgp低的族群的hp wt較高，drat qsec較低；

而disp低的族群的drat qsec較高，hp wt較低。此結果可拿來跟最後multiple regression的結果做比對

MANOVA

= vs

= am

= gear

= carb

= cyl

= cluster1

= cluster2

設 = 0.1

**X** = **μ** + + i=1,2

目的 : 希望測試燃油效率(mpg)與 排氣量(disp)是否與vs am gear carb cyl之間有因果關係。

(1) FOR vs:

(1) H0: ; Ha: at least one

F = 0 ; p-value = 1 > 0.1 => not reject H0 at = 0.1

沒有充分證據顯示vs和cluster有顯著關係

(2) FOR am:

(2) H0’: ; Ha’: at least one

F = 7.42 ; p-value = 0.0106 < 0.1 => reject H0’ at = 0.1

有充分證據顯示am和cluster有顯著關係

(3) FOR gear:

(3) H0’’: ; Ha’’: at least one

F = 4.03 ; p-value = 0.0537 < 0.1 => reject H0’’ at = 0.1

有充分證據顯示gear和cluster有顯著關係

(4) FOR carb:

(3) H0’’’: ; Ha’’’: at least one

F = 0.05 ; p-value = 0.8308 > 0.1 => do not reject H0’’’ at = 0.1

沒有充分證據顯示carb和cluster有顯著關係

(5) FOR cyl:

(3) H0’’’’: ; Ha’’’’: at least one

F = 0.98 ; p-value = 0.3303 > 0.1 => do not reject H0’’’’ at = 0.1

沒有充分證據顯示cyl和cluster有顯著關係

結論 : 故知只有am和gear與mpg、disp有因果關係

Multivariate Multiple regression

目的:希望測試燃油效率(mpg)與排氣量(disp)是否與qesc、hp、drat、wt、am、 gear有關

= mpg

= disp

= qsec

= hp

= drat

= wt

= am

= gear

設

(3)

(i) Model1

，0.8329

， 0.9119

在的情況下進行backward selection

最後剩下的變數為model2

Model2

，0.8497

， 0.9111

For mpg Model2

Test the whole model

vs

= 0.15 ; p-value < 0.001 → reject at , the model is significant.

Test for individual independent variable

vs

= 0.61 ; p-value = 0.0002 < 0.1 → reject at , is significant.

vs

= 0.48 ; p-value < 0.0001 < 0.1 → reject at , is significant.

vs

= 0.87; p-value = 0.0467 < 0.1 → reject at , is significant.

**We remove nothing out of model**

For disp Model2

Test the whole model

vs

= 0.09 ; p-value < 0.001 → reject at , the model is significant.

Test for individual independent variable

vs

= 0.85 ; p-value = 0.0384 < 0.1 → reject at , is significant.

vs

= 0.87 ; p-value =0.0556 < 0.1 → reject at , is significant.

vs

= 0.55; p-value < 0.0001 < 0.1→ reject at , is significant.

vs

= 0.71; p-value = 0.0026 < 0.1→ reject at , is significant.

**We remove nothing out of model**

又因model 2的R平方雖與model1相去不遠，但模型精簡了許多

∴the model we suggest is

結論:

For mpg

燃油效率(mpg)和qsec、wt、am有關

從靜止狀態行駛 1/4 英里的最快時間（以秒為單位）(qsec)每上升一秒，平均燃油效率會上升1.23英里/加侖

車的重量(wt)每上升1 (1000 lbs)，平均燃油效率會減少3.92英里/加侖

若車是手排車(am=1)，平均燃油效率會上升2.94英里/加侖

For disp

排氣量(disp)和qsec、hp、wt、gear有關

從靜止狀態行駛 1/4 英里的最快時間（以秒為單位）(qsec) 每上升一秒，平均排氣量會下降14.77立方英吋

總馬力(hp)每上升一功率，平均排氣量會上升0.47立方英吋

車的重量(wt)每上升1(1000 lbs)，平均排氣量會上升66.86立方英吋

前進檔的數量(gear)每多一個，平均排氣量會下降43.74立方英吋

五、綜合結論及建議

根據cluster結果，mpg低的族群的hp wt較高，drat qsec較低；

而disp低的族群的drat qsec較高，hp wt較低。

根據multiple regression的結果，可知當wt越高，mpg越低(係數為負) ；qsec越低，mpg越低(係數為正)，此符合cluster結果

當qsec越高，disp越低(係數為負) ； hp、wt越低disp越低(係數為正)，此也符合cluster結果

若想知道在multiple regression裡不顯著的連續型變數()的特徵，也可看cluster的結果有初步的推斷。

可惜的是這筆資料不適合用PCA、Factor analysis進行維度刪減，試了將用PCA後，發現她只有一個PCA顯著，並且只有的PCA1係數為0.99，其他變數皆小於0.01，不太能代表資料。還好最後我放棄了PCA與FA才能得出如此好的結果

六、參考文獻

R的內建資料框架：mtcars

<https://2formosa.blogspot.com/2017/10/R-builtin-dataframe-mtcars.html>

SAS output

canonical





Cluster



MANOVA







Multiple regression

For mpg

原本的模型



經backward以後的模型



For disp

原本的模型



經backward以後的模型



最終回歸模型的各種統計量

For mpg for disp

