酒精是否有助於提升學生成績

第七組

組員：

經濟四 林文硯

經濟四 盧柏任

哲學四 鍾宜澄

財金三 張元睿

1. 研究動機

2016年的一則新聞指出，有研究團隊表明酒精能改善短期記憶力，在這個酒精在大學校園泛濫的21世紀，此研究變得格外有趣。

以中正校園而言，便宜酒精的取得非常便利，但是酒精的攝取真的會讓同學的記憶力變好嗎？

我們想透過學生的背景、學習與酒精攝取資料來觀察並驗證此現象。

* 研究顯示飲酒的人在重複詞彙學習任務方面能力較強，甚至喝的量越多，效果會越明顯。

主要的解釋為酒精會阻礙大腦接收新資訊，因此飲酒前獲得的短暫記憶將會被轉入長期記憶區。

1. 變數及樣本
2. 資料背景

我們從Kaggle的網站下載一份有關學生酒精攝取量的資料，這份資料主要說明一些外在環境及社會表徵對他們三次葡萄牙文考試成績的影響。

作者針對在義大利的兩所高中，年齡分布於15~22歲的649名學生進行研究，其蒐集的資料為三類，分別為社經背景、生活狀況和學習狀況，變數總共30種，其中，作者蒐集了兩種成績，分別是數學和葡萄牙文的成績，本報告僅討論葡萄牙文成績。

1. 資料說明

* 自變數(indep.)：

1.School\*(GP. MS)

2.Sex\*(F. M)

3.Age

4.Address

5.famsize\*(GT3.LE3)

6.pstatus

7.medu\*(1-5)

8.fedu

9.mjob

10.fjob\*(T, H, C, A, O)

21.higher\*(yes or no)

22.internet

23.romantic\*(yes or no)

24.famrel

25.freetime

26.goout

27.dalc\*

28.walc

29.health\*(1 - 5)

30.Absences\*(0 - 93)

11.reason\*(H, R, C, O)

12.guardian\*(M,F,O)

13.traveltime

14.studytime\*(1-10Hr)

15.failures\*(if 1<=n<3, else 4)

16.schoolsup\*(yes or no)

17.famsup

18.paid

19.activities

20.nursery

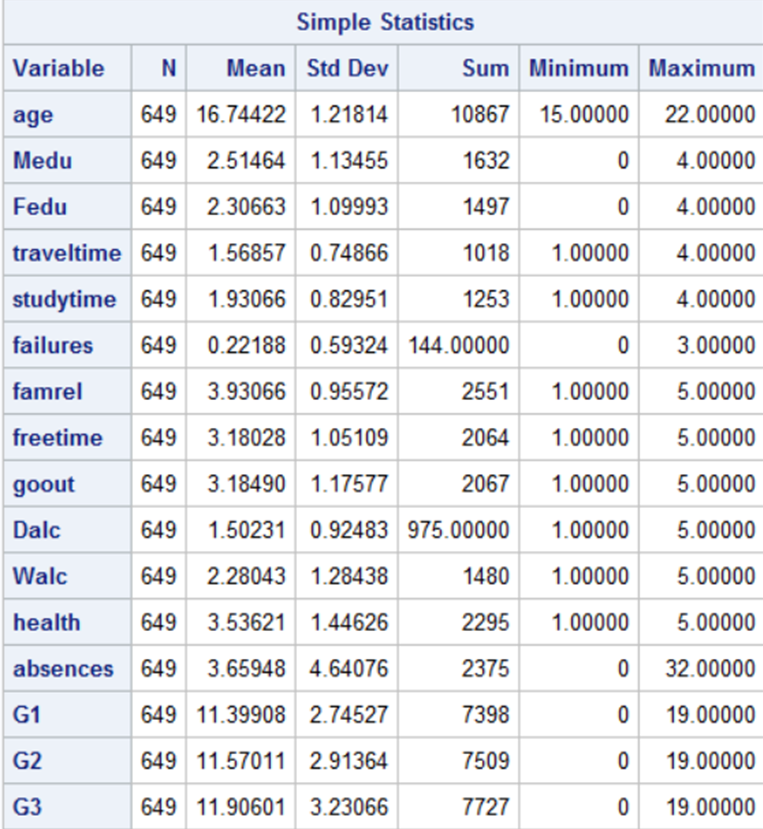
* 應變數(dep.)：

Grade

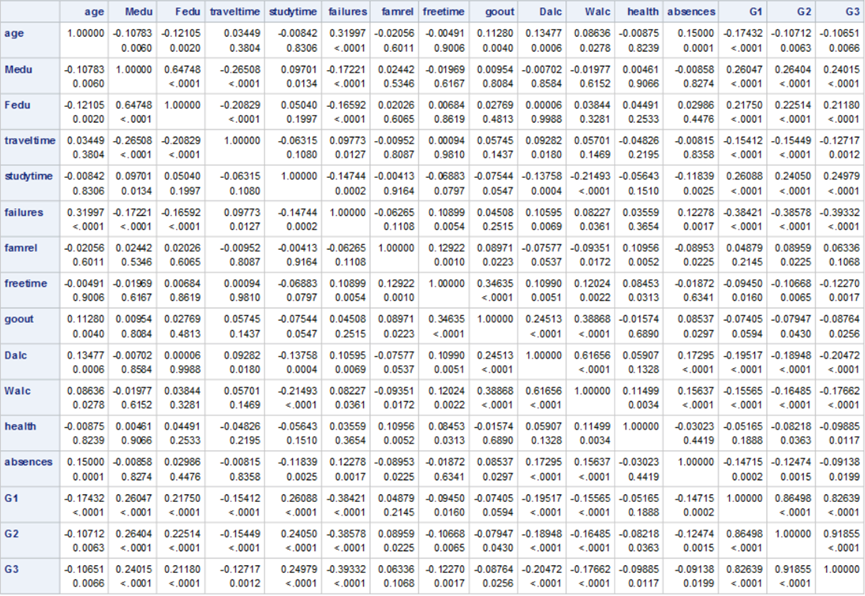
* 類別變數：

school、sex、address、famsize、pstatus、Mjob、Fjob、reason、guardian、schoolsup、famsup、paid、activities、nursery、higher、internet、romantic

* 數值變數：



* 相關係數：



* 相關係數>0.3

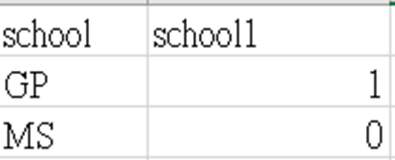
age和failures、Medu和Fedu、Failure和(G1,G2,G3)、feetime和goout、Goout和Walc、Dalc和Walc、G1和G2和G3

1. 流程總覽
2. 資料單鍵編碼及target轉換(平均)
3. 檢查自變數間是否存在共線性問題
4. 篩選變數做迴歸分析
5. target轉換(二元)做決策樹分析
6. 篩選出來的變數(除Dalc、Walc外)的ANOVA分析
7. 主觀判斷影響較深的並固定該變數分析酒精對平均成績的影響(double-sorting portfolio)
8. 結論
9. 資料處理
10. 資料預處理

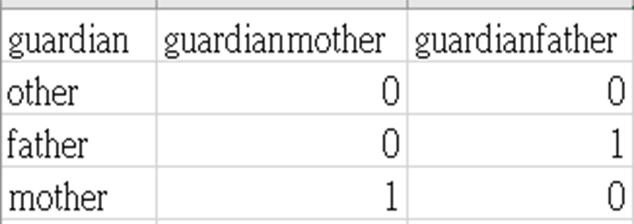
原本變數有30個，但因為裡面有包含類別變數，所以將它做單鍵編碼，如果該分類變數有k個值，選定其中一個值為參照變數，並將其轉換為k-1個虛擬變數，會這樣做的原因是因為我們的分類變數若沒做處理變沒辦法套進模型，而他又是無序的，所以只能改成數個虛擬變數。

Ex：

* school(MS.GP)



* guardian(mother,father,other)



其他變數以此類推，處理後變數從原本的30個變為40個變數。

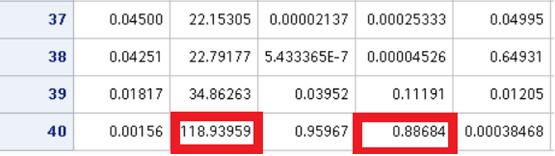
1. 共線性問題

* 以VIF作分析均<10，檢驗沒有很嚴重的共線性問題。





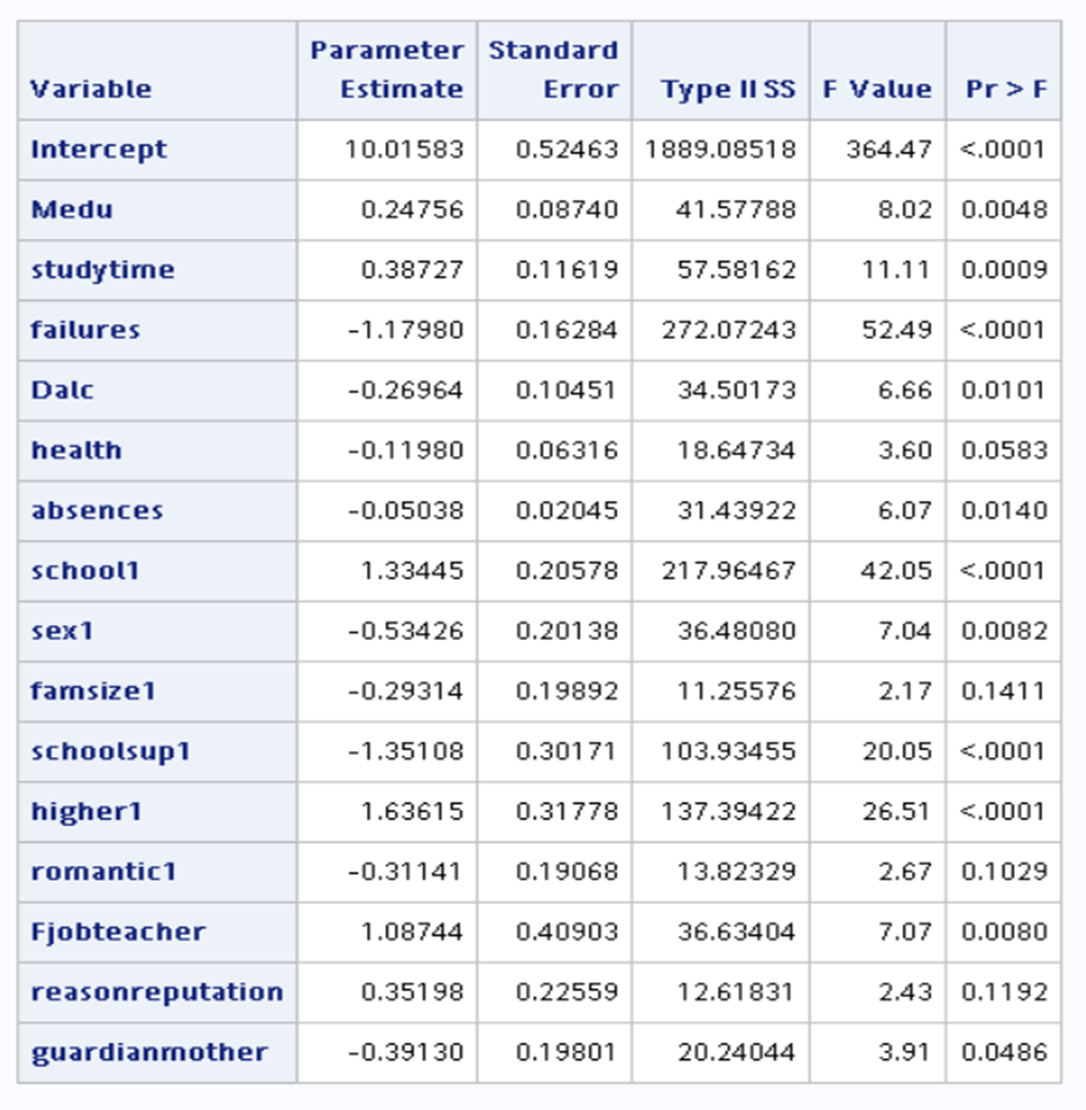
* 圖中截距118.93959>100為condition index，而age的變異數比例則為0.88684，相對其他變數過高，因此我們判斷age存在共線性問題，所以將age刪除。

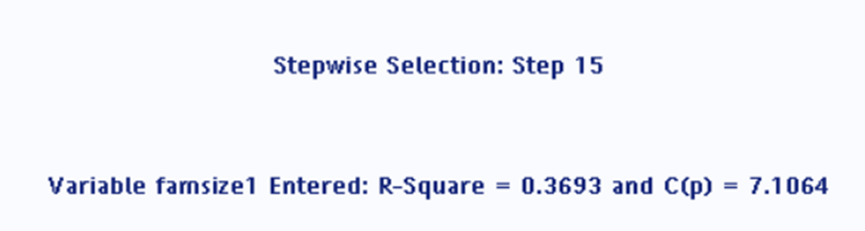


(資料過多，只截取部分結果)

1. 分析模型
2. 選取變數與迴歸分析

將剛剛編碼後的40個變數扣除age後作逐步迴歸，盡量拉高模型的解釋力，最終選取以下變數作為迴歸依據。



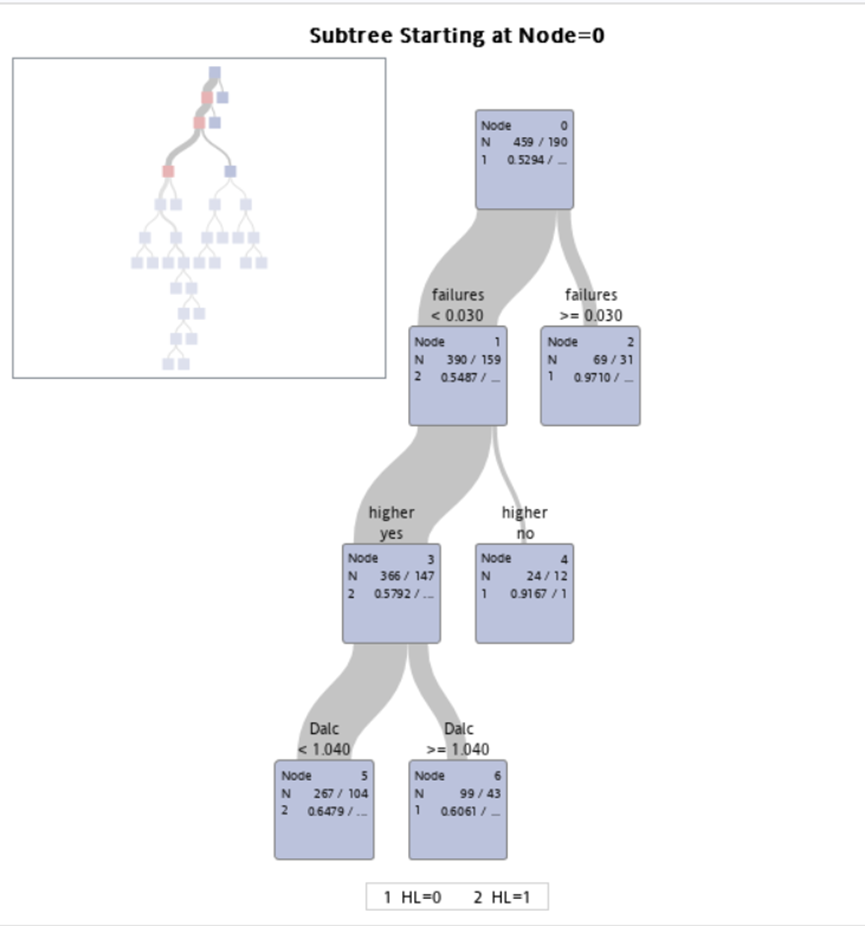


* 迴歸方程式：

Grades=10.01583+0.24756medu+0.38727studytime-1.17980Failures-0.26964Dalc-0.11980Health-0.05038Absence+1.33445School-0.53426Sex-029314Famsize-1.35108Schoolsup+1.63615Higher+0.31141Romantic+1.08744FjobTeacher+0.35198Reasonreputation-0.3913Guardianmother

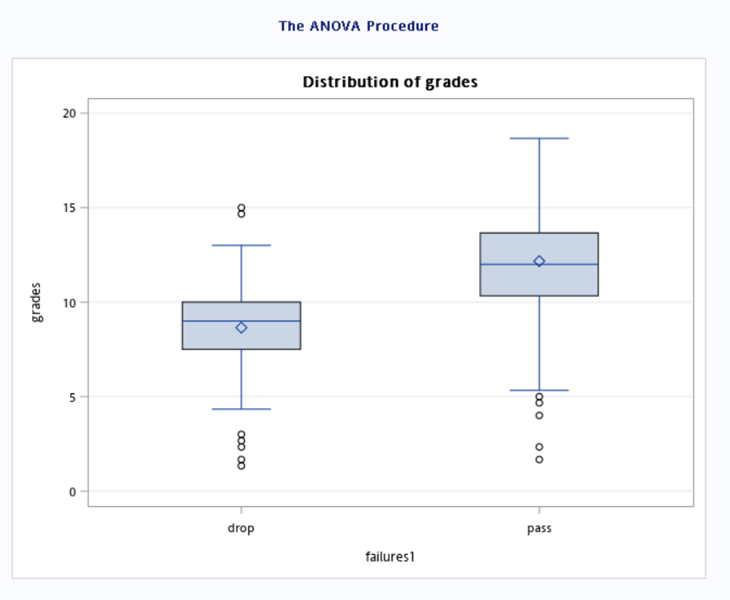
* Grades =(G1+G2+G3)/3

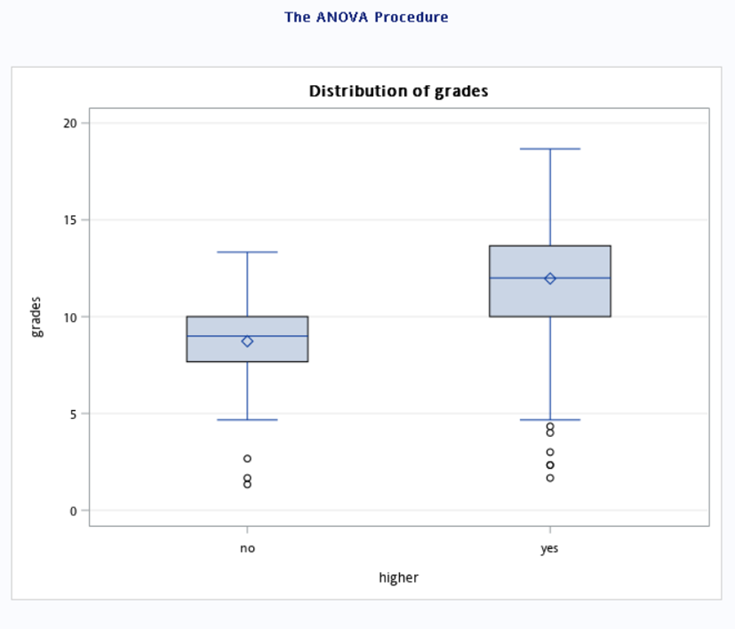
1. 決策樹分析



做決策樹過程中，我們先將平均成績grades依照大於平均與否轉為二元變數HL，從圖中可看出只要曾經被當掉過，幾乎可以確定平均成績低於平均，而其次重要的則是higher(繼續升學)跟Dalc(平日飲酒)，由於此三者皆存在我們的迴歸，特徵重要性又高且出現在決策樹中，因此我們打算固定higher和failure(掛科次數)兩變數去看各自的Dalc及Walc(假日飲酒)對成績的影響，higher保持原本0和1去切分，至於failure，為了不讓切分後的某些樣本變太小，我們也將其作二元分類(有被當過與沒被當過)，進行以下分析。

1. ANOVA分析

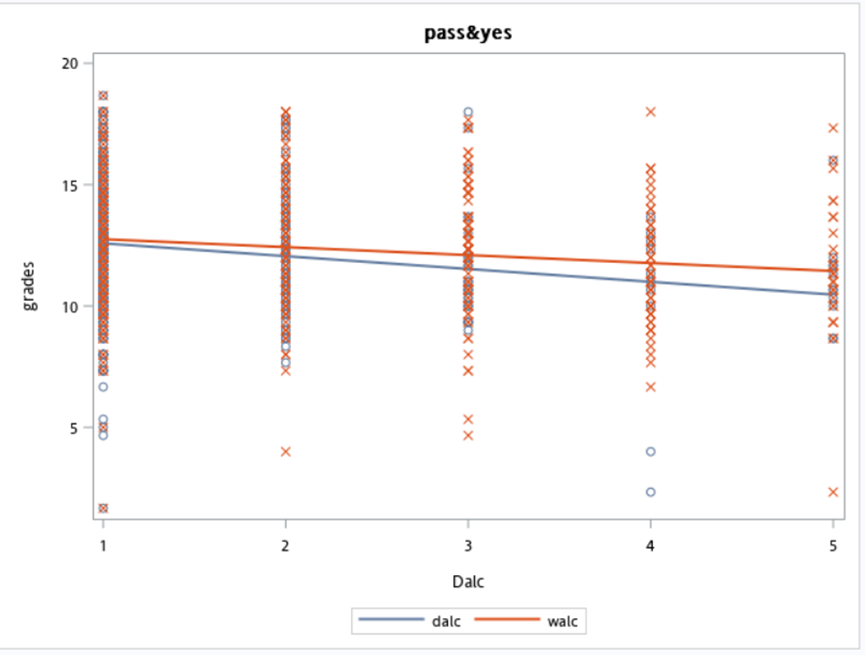


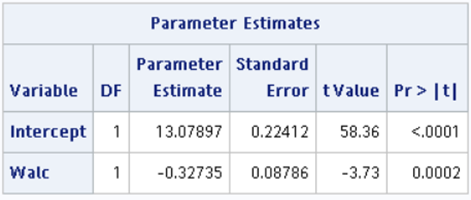
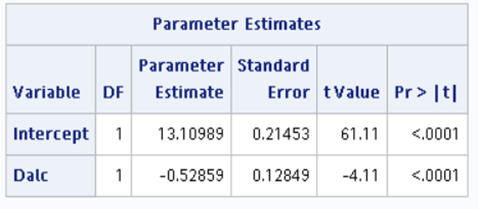


圖中可以看出failures中有被當過的明顯成績低於沒被當過的，而higher則表明願意繼續升學的成績明顯高於不願意的，因此我們固定維度並粗略將其歸為四類並去作各自的Dalc和Walc對grades的迴歸。

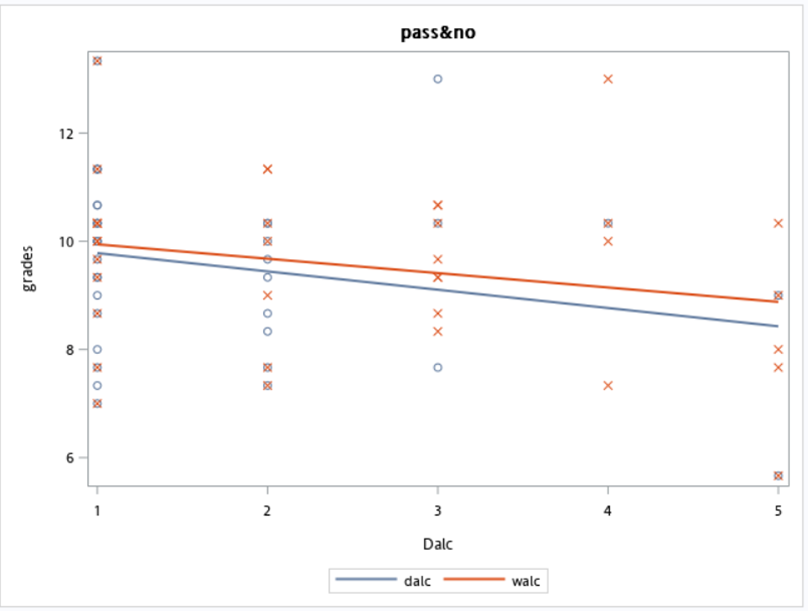
1. 酒精對成績影響

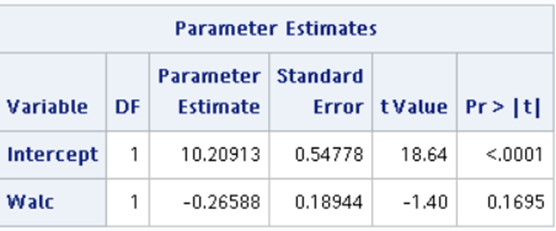
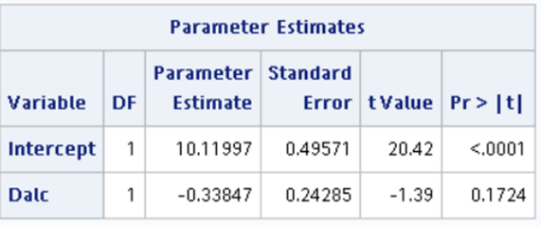
* Pass & Yes(All Pass且繼續升學)



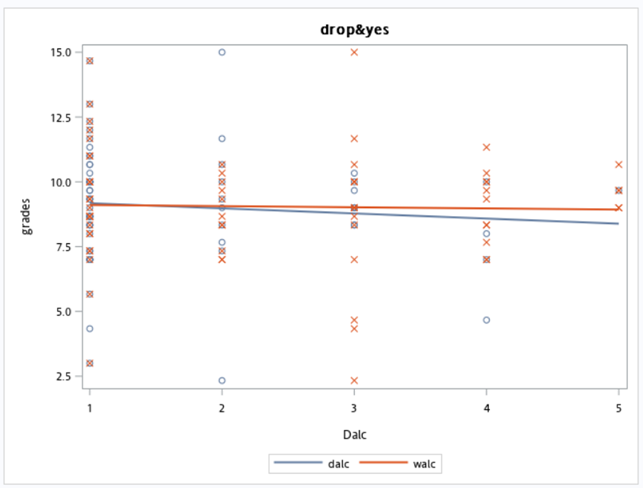


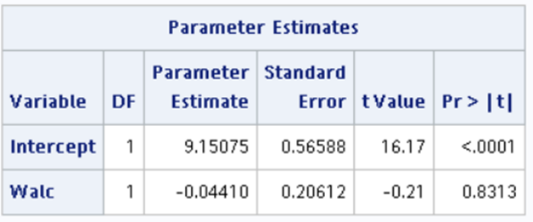
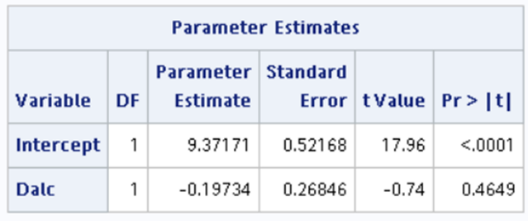
* Pass & No(All Pass且不升學)



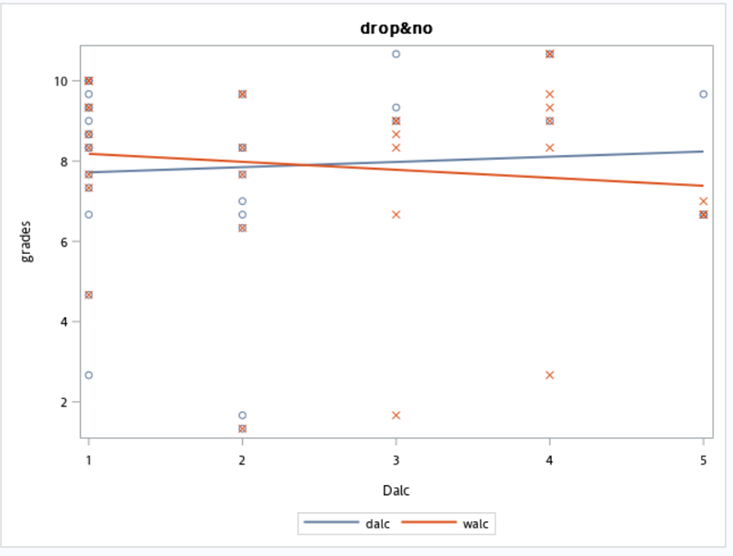


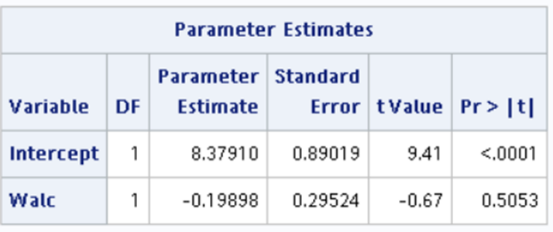
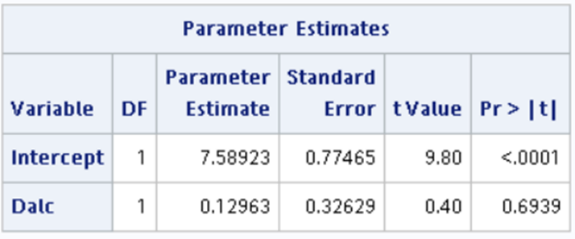
* Drop & Yes(被當過且繼續升學)



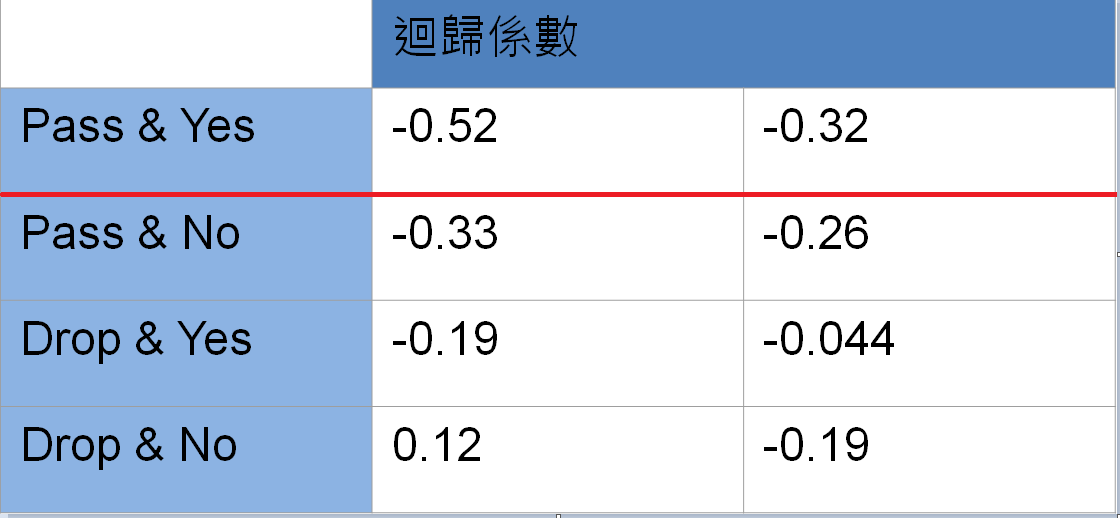


* Drop & No(被當過且不升學)





1. 統計結果



我們以紅線切分，pass & yes組學生成績是否高於平均並不明顯，但從drop和pass & no來看，該組學生成績幾乎都低於平均，再來我們觀察他們的迴歸係數發現pass & yes組相較於成績明顯都不好的學生那幾組迴歸係數較大。

因此我們認為對於成績好的人喝酒影響較大，對於成績本身就差的人，喝酒並無顯著影響。

1. 結論

大體而言，酒精和成績資料上無法說明兩者如同報導所提到的「酒精增強記憶力」，甚至兩者間呈現些微負相關，即喝酒會讓成績下降，分析原因可能是：

* 實驗流程不同
* 取樣方法不同

1. 參考資料

* Improved memory for information learnt before alcohol use in social drinkers tested in a naturalistic setting University of Exeter
* P. Cortez and A. Silva. Using Data Mining to Predict Secondary School Student Performance. In A. Brito and J. Teixeira Eds., Proceedings of 5th Future Business Technology Conference