HW3 RE6091020 汪玄同

11.3

一. 資料描述

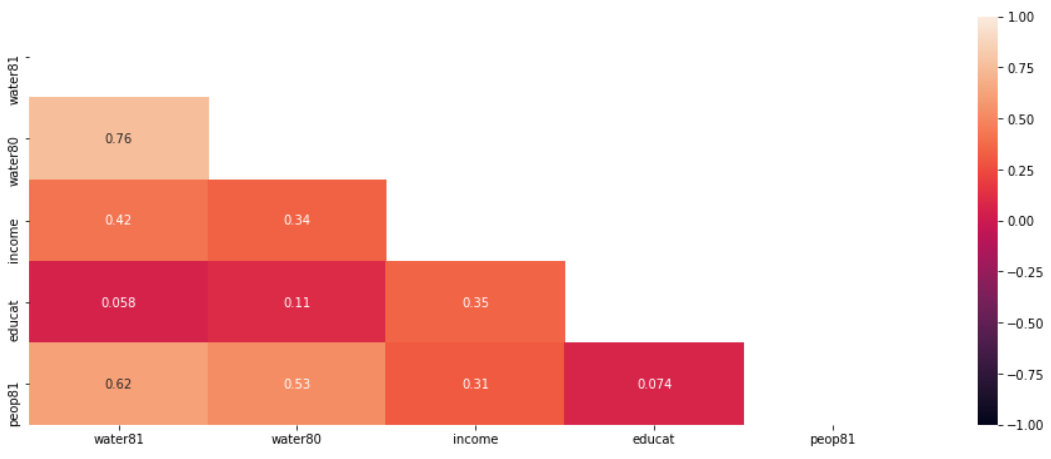
在1970年代後期，新罕布什爾州康科德市嚴重缺水之後，1980年開始採取節水措施。Hamilton收集496個住戶的以下變量信息:

|  |  |
| --- | --- |
| 變數 | 說明 |
| water81 | 1981年家庭用水量(以立方英尺為單位) |
| water80 | 1980年家庭用水量(以立方英尺為單位) |
| income | 1981年家庭收入(1000元為單位) |
| educat | 戶主的教育(以年為單位) |
| peop81 | 1981年夏季居住在家庭中的人數 |
| retired | 戶主退休為1，反之為0 |

目標為用water81為響應變數，其它五項變數為解釋變數建立回歸模型，得到五項變數的係數用來解釋哪些變量會對1981年的用水量造成影響，並且解釋是怎樣的影響，例如解釋變數增加一會增加或減少響應變數多少。

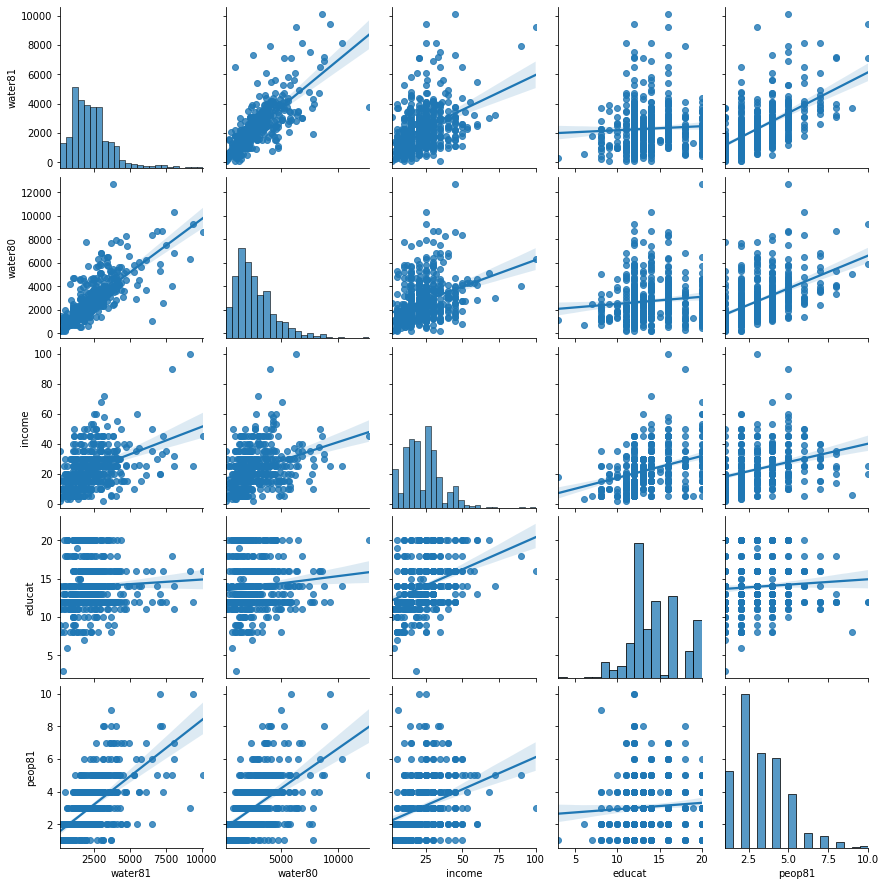
二. 資料探勘

6個解釋變數有5個為連續或離散變數1個為類別變數，下表為5個連續變數對響應變數的相關係數以及解釋變數彼此之間的相關係數熱點圖:

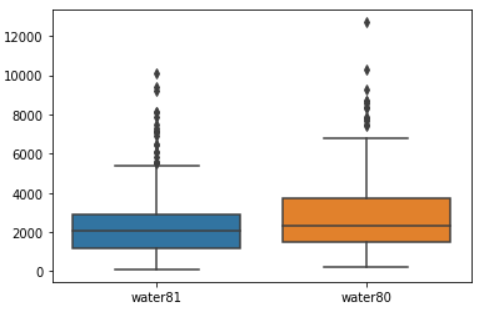
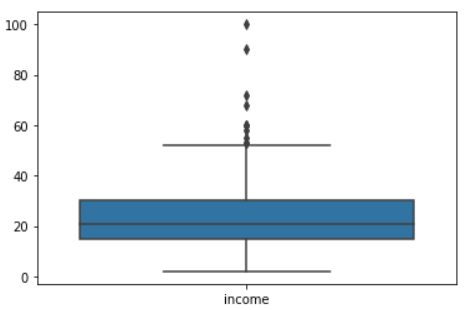
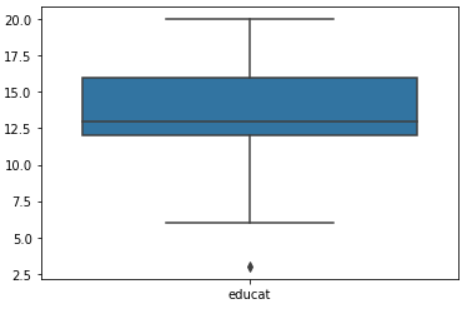
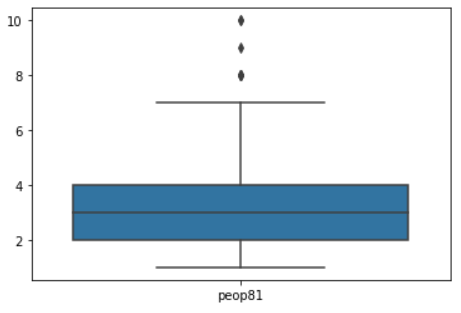


從圖中可以看出每個解釋變數跟81年用水量之間都是正相關，其中教育程度跟用水量關係最小，跟前一年的用水量關係最大。

下圖為scatter matrix:



從scatter matrix中可以看到有一筆資料的80年用水量特別高，兩筆資料的收入特別高。

從盒鬚圖可以看出81年用水量和80年用水量中位數很接近，和其他變數的中位數相差很遠，而因為這些變數尺度相差很大導致數值差距相當大，可能需要考慮做log或取開根號或box cox轉換。

三. 模型建立

先用原資料建立線性回歸，

接下來對模型做常態性檢定，p-value為

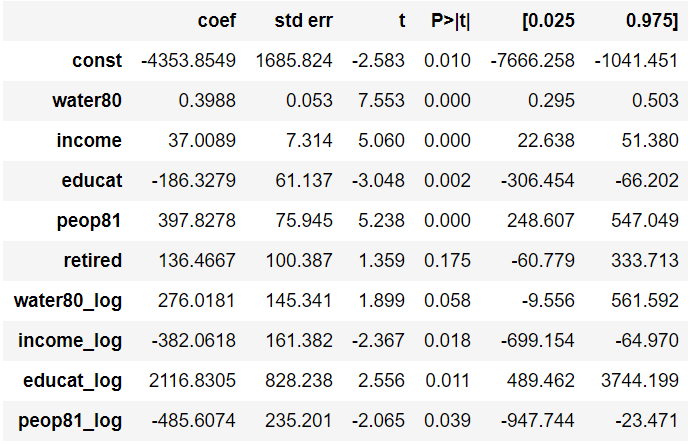
拒絕常態假設，對資料做轉換或增加次方項及交互作用:

由於對資料做各種轉換依然無法通過殘差常態檢定，決定增加次方項或交互作用。

最後結果增加log項的殘差常態性假設p-value為

為增加較多的p-value故使用增加log項的模型為最終模型。

模型係數為:



四. 結論

從模型中的變數係數可以看出沒有經過log轉換的部分只有教育程度和用水量成負相關其他皆成正相關，而從t test的p-value可以看出沒有經過log轉換的變數的係數都顯著不為0經過log轉換的只有教育程度顯著不為0，但因為模型殘差未通過常態假設，所以t test結果可能無法使用。

12.6

一. 資料描述

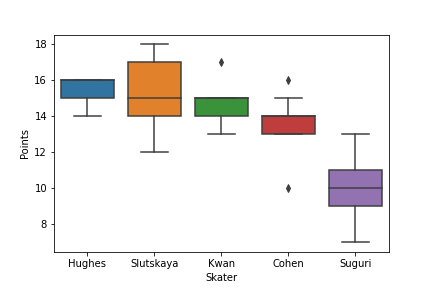
此資料取自2001年的奧委會，包含九名評委對五位選手的技術和表現得分，所以共有45筆資料，每筆資料有四個變數，如下表所示:

|  |  |
| --- | --- |
| 變數 | 說明 |
| Technique | 選手技術得分 |
| Presentation | 選手表現得分 |
| Judge | 九位裁判，以1~9紀錄 |
| Skater | 五名選手，分別為Hughes, Slutskaya, Kwan, Cohen, Suguri |

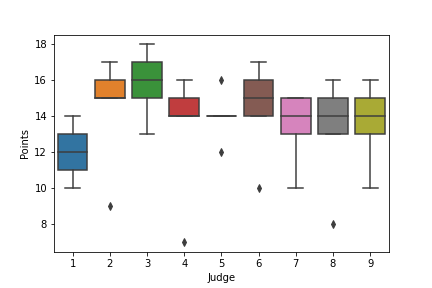
目標為判斷不同選手以及不同裁判的評分一致性是否不同，也就是每位裁判的評分會否因為選手不同而有所不同，以此判斷裁判有沒有偏袒某位選手。

二. 資料探勘

目標的響應變數為技術得分和表現得分的加總，所以將變數Technique和Presentation加總改為變數Points，再用變數Points做不同選手底下的盒鬚圖如下:



假設每位評審的評分不受是哪位參賽選手影響，也就是評審跟選手兩個因子彼此獨立的話，那盒鬚圖表現出來的分數應該要是平移的，標準差應該要盡量相同，但途中可以看出選手Slutskaya和Suguri的盒子標準差與其他三位不太一樣，且Kwan Skater和Cohen兩位選手有特別高和特別低的分數。

在畫出以不同評審為因子的盒鬚圖如下:

可以看出第五位評審畔的分數有一個特別高一個特別低，其他幾乎都相同，這可能是導致上一張盒鬚圖標準差不同的原因。

三. 分析方法

資料有兩個因子分別控制分數，分別為評審和選手，使用two-way anova建立模型:

由於每一個評審只會對一位選手打一次分數，在一個成對因子資料內只會有一筆資料，所以不考慮交互項。

1.常態性:

使用Shapiro-Wilk檢定殘差服從常態，假設如下

統計量與P-value為0.97, 0.31不拒絕殘差服從常態。

2.同質性:

使用Levene’s test檢查不同選手是否符合同質性，假設如下

統計量與P-value為1.05, 0.39不拒絕殘差具同質性假設。

在檢查不同裁判間是否符合同質性

統計量與P-value為0.57, 0.78不拒絕殘差具同質性假設。

3.

P-value為拒絕虛無假設。

四. 結論

每位評審的平均給分有顯著不同，但因為模型通過同質性檢測，所以每位評審給分的變異數沒有相差很大，平均給分不同有可能是評審個人標準不同，無法從這裡看出有沒有偏袒某位選手，建議可以做群體間個別的變異數檢定，看變異數是否相同。