Homework 5--R computing for Business Data Analytics

好棒棒:葉早彬、陳威宇、劉瑞祥

Q1. (10%) Import the library AER in R, and attach the data set CPS1988 (like what you did for HW4).

Focus on four variables – wage, education, experience, and ethnicity (African-American versus Caucasian).
Your job is to finish the following tasks.

(a) Run the linear regression model below (using lm( )) and save the model as object “CPS_lm”.

Note that the ethnicity is a categorical/dummy variable.

```
library("AER")
data("CPS1988")
attach(CPS1988)
experience2=experience*experience
CPS_lm=lm(log(wage)~experience+education+experience2+as.factor(ethnicity))
```

(b) Run the same model again using glm(….., family= “gaussian”). Are the results different from the results from part (a)? If not, why is that?

```
CPS_glm=glm(log(wage)~experience+education+experience2+as.factor(ethnicity), family="gauss
```

"gaussian","gaussian"的假設與lm的假設一樣,所以跑出來的結果也會一樣。

(c) Estimate the same model using quantile regression (section 7.5).Similar to what I do in 7.5, run the model from 5% quantile to 95% quantile, and VISUALIZE the impact of each variable on log(wage) across the whole range of quantiles. Explain what you see.

```
library(quantreg)
cps_f=log(wage)~experience+education+experience2+as.factor(ethnicity)
cps_rqbig=rq(cps_f, tau=seq(0.05, 0.95, 0.05), data=CPS1988)
```

```
cps_rqbigs=summary(cps_rqbig)
plot(cps_rqbigs)
```

Intercept

Intercept與log(wage)的區間有正向的線性關係,wage越大,基本的Intercept越大

experience

experience與log(wage)的區間有負向的線性關係,也就是說wage越大,受experience的影響越小

education

education對後20%的wage,幫助不大,對於位在20%$_{\text{60\%的wage來說}}$,wage受教育的影響越來越大,超過60%後,wage受education的影響差不多。

$experience^2$

$experience^2$ 與log(wage)的區間有正向的線性關係,wage越大,受\$experience$^{\text{2\$的影響越大}}$

ethnicity

ethnicity對log(wage)的影響,在後15%的影響是,wage受ethnicity的影響越來越小,之後影響會越來越大

Q2. (40%) Hand in a short proposal (1.5 lines spacing & LESS than or equal to two A4 pages) for your final term project. At a minimum the proposal should

1) motivate the research question, 2) explain the analysis method to use, 3) briefly describe your data (you do not necessarily need data if you decide to do some theoretical modeling work), and 4) clearly state what the objective of the project will be.

1. motivate the research question

本組想要做的主要是針對線上廣告(online advertising)，根據不同的使用者看到廣告的模式，包含不同的App、不同款式的手機、時間等等變數，來影響使用者是否會點擊廣告。而Click Through Rate(CTR)這是計算廣告點擊的方法之一，其計算方式為使用者看到該類型的廣告總數分之真正點擊廣告次數，本組希望能夠透過資料統計分析來預測及找出何種或那些變數會影響使用者決策判斷最顯著。

2. explain the analysis method to use

這次預測主要因為資料的型態多半是名目變數(Nominal Variable)，因此想採取的方法為決策樹(Decision Tree)，利用計算出每個變數Information Gain的數值來建立決策樹模型，此數值主要是在經過某變數分類前的資料量減去分類後的資料量的多寡，代表著將原本的資料集分散成較小的子集合，而每一個子集合的熵(Entropy, 資料亂度)都會降低，簡單來說就是Information Gain數值越高的變數，越能將資料集清楚的分群到各個子集合，讓原始資料的群集數(一筆資料一個群集)能夠大幅下降，這分類後的子集合群間相似度較低。

3. briefly describe your data (you do not necessarily need data if

you decide to do some theoretical modeling work

本資料是由Kaggle平台上的Data Mining競賽所獲得，是Avazu公司所提供，訓練資料集是10天的CTR資料，主要input包含id,hour,banner_pos等23個變數，而最後想判斷的結果也就是output為click/non-click
以下是資料格式與描述
資料名稱：描述(真實資料範例)
id: ad identifier (1.00001 E+18, 1.00004E+18...)
click: 0/1 for non-click/click (0,1)
hour: format is YYMMDDHH, so 14091123 means 23:00 on Sept. 11, 2014 UTC.(14102100, 14102102, 14102104...)
C1 -- anonymized categorical variable (1005,1002....)
banner_pos:banner position(0,1)
site_id: site identrifier(1fbe01f1, d9750ee7...)
site_domain:site domain(f3845767, c7ca3108...)
site_category:site category(28905ebd, f028772b...)
app_id:app identifier(ecad2386, 98fed791...)
app_domain: app domain(7801e8d9, d9b5648e...)
app_category: app category(07d7df22, cef3e649...)
device_id:device identifier(a99f214a, fb3c543...)
device_ip:device ip(ddd2926e, b3cf8def...)
device_model:device model(44956a24, c26c53cf...)
device_type:device type(0,1,4,5)
device_conn_type: device connection type(0,2,3,5)
C14-C21 -- anonymized categorical variables(from C14 to C21: 15706, 320, 50, 1722,0,35,-1,79)

4. clearly state what the objective of the project will be.

本預測主要是透過決策樹辨別出某些類別能夠將資料群集有效分類，辨別出那些變數會影響結果，及這群變數應該設定成怎樣才能有較好的CTR，讓廣告主可根據此決策數的pattern來設計廣告，提高廣告點擊率，達到廣告效益。

Q3. (60%) Did you attend Prof. Chia-Yen Lee’s talk on November 25?

yes