Statistic\_Project1

Syuan-Bo,Huang

2022/4/7

Table of Contents

knitr::opts\_chunk$set(echo = TRUE)

## 資料簡介

#### **Title**：the amount of page views for the top 1,000 websites on the Internet as of 2011

Focus on 5 columns

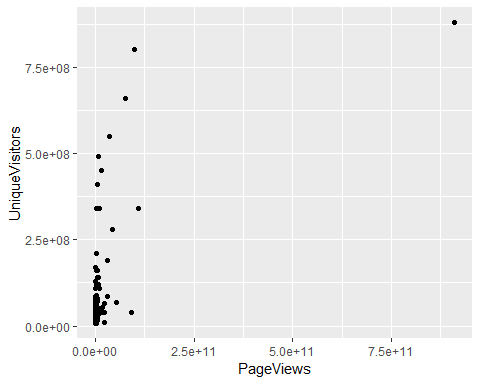
1. **Rank**: 網站排名
2. **PageViews**: 該網站一年總瀏覽次數(我們要預測的資料)
3. **UniqueVisitors**: 該網站一年不重複人次的瀏覽次數
4. **HasAdvertising**: 該網站有無廣告
5. **IsEnglish**: 該網站的主要語言是否為英文

top.1000.sites <- read.csv('https://raw.githubusercontent.com/johnmyleswhite/ML\_for\_Hackers/master/05-Regression/data/top\_1000\_sites.tsv' ,sep = '\t',stringsAsFactors = FALSE)  
head(top.1000.sites[,c(1,4,6,7,8)],n=10)

## Rank UniqueVisitors PageViews HasAdvertising InEnglish  
## 1 1 880000000 9.1e+11 Yes Yes  
## 2 2 800000000 1.0e+11 Yes Yes  
## 3 3 660000000 7.7e+10 Yes Yes  
## 4 4 550000000 3.6e+10 Yes Yes  
## 5 5 490000000 7.0e+09 No Yes  
## 6 6 450000000 1.5e+10 Yes Yes  
## 7 7 410000000 5.4e+09 Yes Yes  
## 8 8 340000000 1.1e+11 Yes No  
## 9 9 340000000 1.1e+10 Yes Yes  
## 10 10 340000000 2.7e+09 Yes Yes

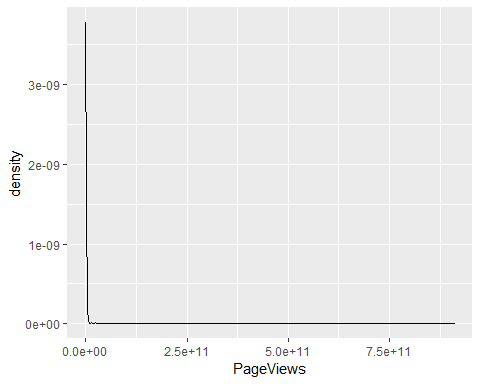
## 分析概念及步驟說明

PageViews是預測的對象 開始做迴歸分析之前，先繪製 **PageViews** 與 **UniqueVistors** 間的**散布圖**，觀察兩者數值資料是否呈現性關係。



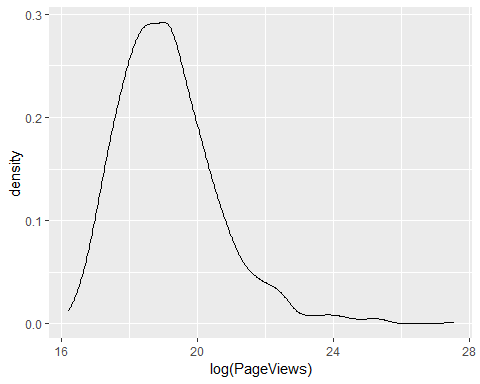
由上圖可發現兩者線性關係薄弱，從資料分布狀況推測為圖面比例尺問題，導致大部分資料集中於圖面左下角。 接著單獨觀察 **PageViews** 的資料分布情形。

ggplot(top.1000.sites,aes(x=PageViews))+geom\_density()



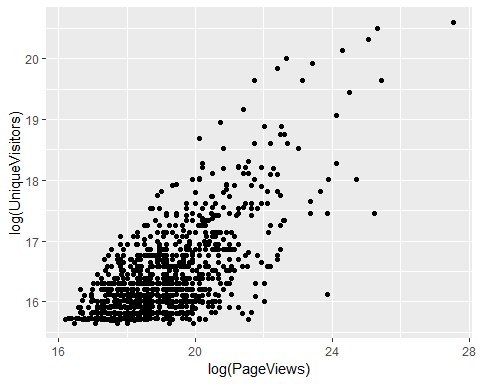
結果呈現出無意義的密度圖，嘗試使用log轉換後的值繪製密度圖。

ggplot(top.1000.sites,aes(x=log(PageViews)))+geom\_density()



呈現密度圖較合乎邏輯，因此接下來採用log轉換後的PageViews和UniqueVisitors進行分析。 重新製作兩者間log轉換後的散佈圖。

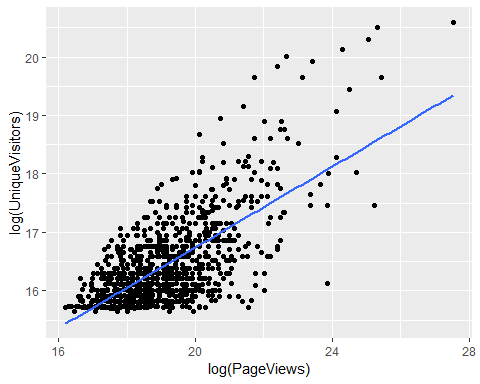
ggplot(top.1000.sites,aes(x=log(PageViews),y=log(UniqueVisitors)))+geom\_point()



其散佈結果呈現線性關係，使用method ’lm’去描繪迴歸線

ggplot(top.1000.sites, aes(x =log(PageViews),y=log(UniqueVisitors))) +geom\_point()+geom\_smooth(method = 'lm', se = FALSE)

## `geom\_smooth()` using formula 'y ~ x'



## 分析結果

使用summary function做迴歸分析計算並總結分析結果

lm.fit <- lm(log(PageViews) ~ log(UniqueVisitors),  
data = top.1000.sites)  
summary(lm.fit)

##   
## Call:  
## lm(formula = log(PageViews) ~ log(UniqueVisitors), data = top.1000.sites)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.1825 -0.7986 -0.0741 0.6467 5.1549   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -2.83441 0.75201 -3.769 0.000173 \*\*\*  
## log(UniqueVisitors) 1.33628 0.04568 29.251 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.084 on 998 degrees of freedom  
## Multiple R-squared: 0.4616, Adjusted R-squared: 0.4611   
## F-statistic: 855.6 on 1 and 998 DF, p-value: < 2.2e-16

log(UniqueVisitor)的t-value為29.251(coef/std)，遠大於2，且p-value為2\*10\*\*-16小於0.001，不大於0.05，因此log(UniqeiVisitors)的係數非0，表示log(UniqueVisitors)和log(PageViews)兩者間呈線性關係。 兩者的R square值為0.4616，表示這個線性迴歸模型對這筆資料有46%的解釋性，表現還不錯。 最後F-statistic的值為855.6，遠大於1，表示拒絕係數等於0的虛無假設，代表至少存在一個不為0的係數。

接者觀察加入HasAdvertising和IsEnglish後的關係

lm.fit <- lm(log(PageViews) ~ HasAdvertising + log(UniqueVisitors) + InEnglish,  
data = top.1000.sites)  
summary(lm.fit)

##   
## Call:  
## lm(formula = log(PageViews) ~ HasAdvertising + log(UniqueVisitors) +   
## InEnglish, data = top.1000.sites)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.4283 -0.7685 -0.0632 0.6298 5.4133   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -1.94502 1.14777 -1.695 0.09046 .   
## HasAdvertisingYes 0.30595 0.09170 3.336 0.00088 \*\*\*  
## log(UniqueVisitors) 1.26507 0.07053 17.936 < 2e-16 \*\*\*  
## InEnglishNo 0.83468 0.20860 4.001 6.77e-05 \*\*\*  
## InEnglishYes -0.16913 0.20424 -0.828 0.40780   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.067 on 995 degrees of freedom  
## Multiple R-squared: 0.4798, Adjusted R-squared: 0.4777   
## F-statistic: 229.4 on 4 and 995 DF, p-value: < 2.2e-16

lm.fit <- lm(log(PageViews) ~ HasAdvertising,  
data = top.1000.sites)  
summary(lm.fit)$r.squared

## [1] 0.01073766

lm.fit <- lm(log(PageViews) ~ log(UniqueVisitors),  
data = top.1000.sites)  
summary(lm.fit)$r.squared

## [1] 0.4615985

lm.fit <- lm(log(PageViews) ~ InEnglish,  
data = top.1000.sites)  
summary(lm.fit)$r.squared

## [1] 0.3043425

## Referance

* Drew Conway and John Myles White (2012) Machine Learning for Hackers, 1st Edition, O’Reilly, Sebastopol, CA.