PTT Analysis of Entrance Exam Scores in Taiwan

Christine P. Chai

October 20, 2019

I just started the template.

Executive Summary

Write something here

Introduction

I am curious about the relationship between the high school entrance exam score percentiles and the college entrance scores in Taiwan. It seems obvious that a greater percentage of students from top high schools get admitted to prestigious universities[1].

However, the high school entrance exam is much easier than the college entrance exam, so some people studied little in middle school and was able to get into a good high school. Then some of these people kept studying little and ended up with a bad score on the college entrance exam. On the other hand, I have also seen some students from mediocre high schools worked very hard during the three years, and eventually earned a stellar score in the college exam.

Therefore, I decided to gather data and create my own analysis.

Background

In Taiwan, the high school entrance exam score percentiles are between 1% and 99%, and people often refer to the percentile rank (PR value) as from 1 to 99. This scoring system existed from 2001 to 2013[2]. The actual exam score ranges were different. For example, the maximum possible score was originally set to 300, but it was increased to 312 in Year 2007. Then the maximum possible score was increased to 412 in Year 2009. Therefore, the percentile ranks (PR values) serves as a normalized tool to compare academic achievents across different years.

The college entrance exams are held twice a year in Taiwan. The first exam, typically held in late January or early February, is called the General Scholastic Ability Test (GSAT)[3]. The second exam is called the Advanced Subjects Test (AST)[4], and it is almost always held on July 1st, 2nd, and 3rd. The GSAT scores are normalized to a range of 0 to 75, regardless of the difficulty level of GSAT each year. On the other hand, the scores of AST can vary widely because each subject is scored separately from 0 to 100. Since the AST scores fluctuate more due to the difficulty level of the exam questions each year, I decided to use the GSAT scores as a benchmark of the college exam scores.

Data Description

It is a challenge to obtain individual pairs of data as a representative sample. Although it is easy to send out a spreadsheet and ask my friends to report their scores anonymously, this approach can result in a large

[1] https://bit.ly/2JSPXKc
[2] https://bit.ly/2JNQaOI
[3] https://bit.ly/2W0fdUq
[4] https://bit.ly/2J7YxoW

selection bias. Many of my friends graduated from the same high school or college as I did, so we are likely to have similar entrance exam scores.

Hence I retrieved data from the SENIORHIGH (high school)[5] discussion section on PTT[6], the largest terminal-based bulletin board in Taiwan. I assume the data to be more representative (than if I had collected on my own) because anyone can get a PTT account and reply to the post. The majority of scores were reported in May 2015, and a few scores were reported in the following month or later.

Unfinished below

In the data *ptt_SENIORHIGH_data.csv*, the main variables are:

- **pttID**: Each person's ID on PTT, which can be anonymous. This column serves as the unique identifier of each person.
- **HighSchool_PR**: Each person's percentile rank (PR value) of the high school entrance exam.
- **CollegeEntrance_Score**: Each person's GSAT score, which is regarded as the person's college entrance score.

Here are some indicators:

- **HS_Inaccurate**: TBD
- **College_Inaccurate**: TBD

Raw data

Showing the first 10 rows of data.

```
# Data Source: ptt - SENIORHIGH
# https://www.ptt.cc/bbs/SENIORHIGH/M.1432729401.A.995.html
# 197 rows of (high school entrance exam percentile, 1st college entrance exam score)

data = read.csv("ptt_SENIORHIGH_data.csv")
names(data)[1] = "pttID"

data[1:10,]
```

```
##           pttID HighSchool_PR CollegeEntrance_Score HS_Inaccurate
## 1    game275415            60                    50             1
## 2      a2654133            60                    52            NA
## 3   cookie20125            99                    72            NA
## 4       heejung            92                    54             1
## 5        shun01            87                    51            NA
## 6     robinyu85            -1                    74            NA
## 7    allengoose            69                    48            NA
## 8  godpatrick11            98                    60            NA
## 9     morgankhs            95                    65            NA
## 10      jazzard            88                    65            NA
##    College_Inaccurate
## 1                  NA
## 2                  NA
## 3                  NA
## 4                  NA
## 5                  NA
## 6                  NA
```

[5] https://www.ptt.cc/bbs/SENIORHIGH/M.1432729401.A.995.html

[6] If you have a PTT account, you can log into the website using a browser. https://iamchucky.github.io/PttChrome/?site=ptt.cc

```
## 7                   NA
## 8                   NA
## 9                   NA
## 10                  NA
```

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

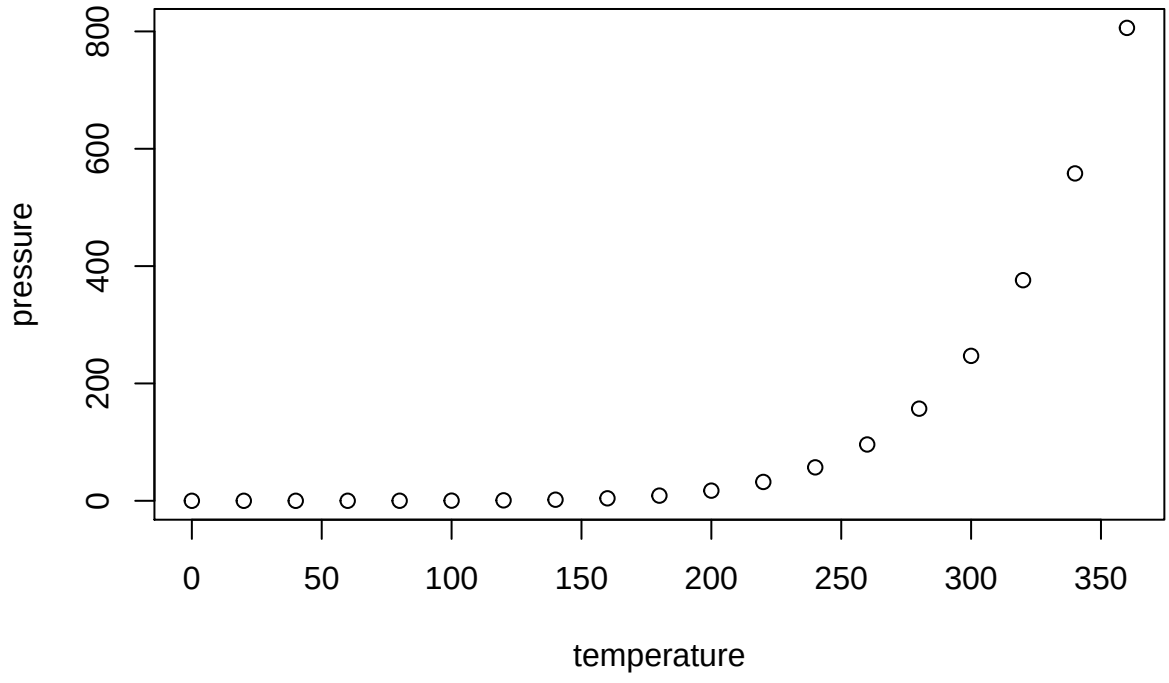
When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
summary(cars)
```

```
##      speed           dist       
##  Min.   : 4.0   Min.   :  2.00  
##  1st Qu.:12.0   1st Qu.: 26.00  
##  Median :15.0   Median : 36.00  
##  Mean   :15.4   Mean   : 42.98  
##  3rd Qu.:19.0   3rd Qu.: 56.00  
##  Max.   :25.0   Max.   :120.00
```

Including Plots

You can also embed plots, for example:



Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that

generated the plot.