# Statistical Analysis of High School and College Entrance Exam Scores in Taiwan with Online Data

A Fully Reproducible Approach with R Markdown

Christine P. Chai
Independent (to be updated)
cpchai21@gmail.com

April 23, 2024

Ongoing work since 2019.

## Executive Summary

In this project, we investigate the relationship between high school and college entrance exam scores in Taiwan using self-reported data online. The goal is to demonstrate reproducible statistical analysis in R Markdown and LaTeX to PDF. We start by exploring the data to formulate the problem statement and identify the appropriate model, which can be an iterative process. We eventually decided on a binary outcome of the college entrance exams, and hence implemented a logistic regression to fit the model. We also validated the model with out-of-sample prediction methods, including cross validation as well as separate training and testing sets. The end-to-end data analysis can be compiled with a single button in RStudio, without copying-and-pasting output from one program to another.

## Disclaimer

The opinions and views expressed in this manuscript are those of the author and do not necessarily state or reflect those of **[Institution Name]**. The author has a statistics and engineering background, without any training in middle school or high school teaching.

## 1 Introduction

The author grew up in Taiwan, so we are 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,[2] so some students studied little in middle school and was able to get into a good high school. Then some of them kept studying little and ended up with a bad score on the college entrance exam. On the other hand, we have also seen some students from non-top-tier high schools studied very hard, and eventually earned a stellar score in the college exam.[3] Therefore, we decided to gather data and create our own analysis.

The Taiwan government publishes aggregated data on high school entrance exam scores[4] and college entrance

[1] https://bit.ly/2JSPXKc
[2] https://www.merit-times.com/NewsPage.aspx?unid=48358
[3] https://www.pttweb.cc/bbs/NTU_CK_CM/M.1235046292.A.41C
[4] https://www.ceec.edu.tw/

exam scores[5] separately, but does not link between the two sets of data by anonymized individuals. (The closest linkage is the college admission numbers for selected high schools.[6]) Therefore, we obtained self-reported data from an existing post[7] in 2015 on PTT, the largest terminal-based bulletin board in Taiwan. Each of the 197 rows of the data contains an anonymous ID, high school entrance exam score (in the form of percentile rank), and college entrance exam score. Since PTT has been advertised to a wide range of college campuses, many college students as of 2015 were using PTT.[8] Hence the PTT data could be assumed to be diverse, i.e., from people with different levels of academic achievements.

The goal of writing this manuscript is to demonstrate how to perform data analysis in `R` using real-life data with a reproducible workflow. We also make this project open-source on GitHub,[9] so that anyone interested can leverage the data analysis framework as well as the statistical methodology. The target audience is people with a basic understanding of statistics, equivalent to taking or having completed Statistics 101. We start from collecting the data, preprocessing the data, exploratory analysis, and finally define a problem statement that can be answered by the data. Then we build the statistical model, validate the results, and provide our interpretations. In this way, readers can follow the step-by-step instructions in this small-scale data science project. Readers will also learn how to handle issues and continue the analysis in a real application problem.

The rest of this manuscript is organized as follows. Chapter 2 gives an overview of the high school and college entrance exams in Taiwan, and Chapter 3 provides the description of our dataset. In Chapter 4, we start the technical content with exploratory data analysis, in order to identify and visualize the distribution of exam scores. Then in Chapter 5 we try the linear regression and explain whether this model is appropriate or not. Chapter 6 further examines the top scorers, because they typically came from prestigious high schools and/or colleges. In Chapter 7, we modify the problem statement to binary outcomes and implement the logistic regression. Then we perform in-sample validation in Chapter 8 and out-of-sample validation in Chapter 9. After the model validation is complete, we use some metrics to evaluate the model performance in Chapter 10. Towards the end, we provide a recap of the project in Chapter 11 and recommend more resources for learning in Chapter 12. Finally, we state some personal remarks in Chapter 13.

# 2 Background

## Unfinished below

Elaborate on the following to enhance the background context:

- College entrance exam GSAT and AST: New rules in 2022 (not related to COVID) due to the 2019 education reform in Taiwan.[10]
- COVID-related rules for students who could not attend the exam.
  Start writing about 2020, 2021, 2022 for both exams.
- If we continue writing this manuscript in 2023, we absolutely need to add the recent changes in 2021 and 2022.

This is the file we want to move content into.

## Unfinished above

We briefly explain the high school and college entrance exams in Taiwan for readers all over the world. Like many other places in Asia, Taiwan used to put heavy emphasis on standardized tests in high school and college

[5] https://cap.rcpet.edu.tw/
[6] https://www.ptt.cc/bbs/juniorhigh/M.1407508506.A.1C9.html
[7] https://www.ptt.cc/bbs/SENIORHIGH/M.1432729401.A.995.html
[8] PTT Screenshot obtained on June 24, 2022: https://imgur.com/Y0pHLiR
[9] https://github.com/star1327p/Exam-Scores-PTT
[10] http://tinyurl.com/3r5zrzhs

admissions.[11] This is different from the US college admission system, which values more of students' overall qualifications such as extracurricular activities and personal qualities, not solely on a single standardized test score.[12]

Nowadays, Taiwan has added some non-exam components for middle and high school students, in order to educate them to be well-rounded.[13] Despite multiple attempts of education reform, the high school entrance exam in Taiwan remains a key part to high school admissions (Chen and Huang, 2017; Chen, 2008). The college admission process in Taiwan is a more complicated system (Liu, 2022). Although Taiwan's college admission process requires students to submit their portfolio as part of their application package,[14] students still need to score high enough in the college entrance exam to pass the initial screening round.[15]

## 2.1 High School Entrance Exam in Taiwan

Between 2001 and 2013, the high school entrance exam in Taiwan was officially called The Basic Competence Test for Junior High School Students (Chou and Ho, 2007).[16] The exam consisted of five subjects (Chinese, English, mathematics, physical sciences, and social sciences). Note that two exams was held each year in 2001-2011, and students could take one or both exams because the difficulty level was about equal.[17] The first exam was during May-June and the second one was in July. Each student could use either set of score (typically the better one) to get their high school placement. The report card contains exam scores for each subject, the total score, and the percentile rank (PR values from 1 to 99) of the total score. The PR value refers to the percentage of the student population you scored higher than. For example, PR 86 means you have a higher score than 86% of the student population in Taiwan. The PR values serve as a normalized tool to compare academic achievements across years.

While the PR values remained consistent, the actual exam score range varied across years. In 2001-2006, each of the five subjects had a maximum score of 60, summing up to a total score of 300. In Year 2007, the independent essay question was added for another 12 points, bringing up the maximum total score to 312 points.[18] The scoring of the high school entrance exam had been controversial and generated much discussion in the news.[19] The main issue was the large score difference between full marks and getting only one question wrong in a test subject.[20] For instance, one could get a 60 with all questions correct in a subject, but only 55 by missing just one question. The five-point deduction was regarded as a harsh penalty, given that the knowledge difference was minimal between a perfect score and a near-perfect score.[21] Then in Year 2009, the maximum possible score was increased from 312 to 412 points.[22] Each subject was augmented from 60 to 80 points, and the independent essay question was still worth 12 points. In the 80-point system, missing one question would result in a two-point loss from the full mark, rather than a substantial five or six points.[23]

In the later stages of this high school entrance exam format, subtle changes were made in 2011 and 2012, until the final year in 2013. The high school entrance exam had used the same questions across the whole country, but in Year 2011, Taipei and its surrounding areas created their own version of entrance exam.[24] The education officials claimed that more granularity was needed to distinguish students' academic ability in such competitive areas.[25] Then in Year 2012, the high school entrance exam was suddenly reduced to one exam per year instead of two,[26] and this continued in Year 2013.[27] In 2012 and 2013, the "second exam"

[11] https://en.wikipedia.org/wiki/Education_in_Taiwan
[12] https://www.usnews.com/education/best-colleges/articles/college-application-process
[13] http://www.ater.org.tw/journal/article/9-6/free/12.pdf
[14] https://www.reallygood.com.tw/newExam/inside?str=FD956A48971EAE124F46CACDAFD17584
[15] https://www.thenewslens.com/article/179752
[16] https://bit.ly/2JNQaOI
[17] https://www.merit-times.com/NewsPage.aspx?unid=120645
[18] https://www.945enet.com.tw/epaper/contents/ha/048/03.htm
[19] https://web.archive.org/web/20090415123425/http://www.libertytimes.com.tw/2007/new/jun/21/today-life1.htm
[20] https://blog.xuite.net/wmhsu/twblog/147553826
[21] https://m.xuite.net/blog/chioufatymjh/twblog/130262492
[22] https://blog.cybertranslator.idv.tw/archives/2384
[23] https://joan044.pixnet.net/blog/post/255140987
[24] https://sites.google.com/site/kleduentrance/01001/07
[25] https://bit.ly/3Zq2c6e
[26] https://m.xuite.net/blog/penghu.dialy/blog/61024095
[27] https://www.ettoday.net/news/20130609/220042.htm

was replaced with a pre-exam admission process: Each middle school could submit formal recommendations for their students to high schools, with a limited quota per high school. Students who got admitted by a high school would not have to take the entrance exam.[28] The controversy was that the changes were implemented quickly, so neither students nor teachers in middle schools were adequately prepared to make optimal choices.[29]

A major reform came in Year 2014, and the high school entrance exam was officially replaced with the Comprehensive Assessment Program (CAP) for Junior High School Students.[30] Instead of using PR values that emphasized the ranking, CAP provides each student with seven letter grades for each subject (C, B, B+, B++, A, A+, A++) and an independent essay score (0-6).[31] Each region in Taiwan sets their own rules for high school admission decisions,[32] and the CAP score is a necessary but not sufficient condition to get into a top high school. For example, Taipei and its surrounding areas award points for completing five semesters of physical education, arts and humanities, and integrative activities (with an average grade of C or better).[33] Tainan (southern part of Taiwan) also requires additional criteria such as physical fitness, volunteering activities, and participation in student organizations.[34]

**Remark**: Since the data were self-reported in 2015, the high school entrance exam scores included only PR values and did not include any letter grades. In the CAP era starting in 2014, the wide variance of high school admission requirements makes it more difficult to compare across students' academic achievements.

# Unfinished below

## 2.2 College Entrance Exam in Taiwan

The college entrance exams in Taiwan are held twice a year by the College Entrance Examination Center (CEEC).[35] The first exam is called the General Scholastic Ability Test (GSAT), and the scores are used for early admission. The second exam is called the Advanced Subjects Test (AST), and the scores are used for regular admission. The college admission rules change slightly every several years, but the format of two entrance exams has been used for 20+ years and is here to stay (Chou, 2015; Chiang, 2022).

### 2.2.1 General Scholastic Ability Test (GSAT)

The GSAT has been in place since 1994, and the exam is typically held in late January or early February.[36] Students can use the GSAT scores and additional criteria (i.e., their portfolio documents) to apply to colleges for early admission (Hsieh, 2019). Starting in 2022, the early admission from GSAT is not as "early" anymore.[37] The timeline of the first round of college admission has been postponed from March to May, leaving high school students little time to prepare for the upcoming second exam in July.[38]

The GSAT consists of five subjects (Chinese, English, mathematics, physical sciences, and social sciences), each of which are graded on a 0 to 15 point scale. 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, we decided to use the GSAT scores as a benchmark of the college exam scores.

The 0 to 15 point scale is calculated in three steps:[39]

[28] https://news.ltn.com.tw/news/life/breakingnews/537370
[29] https://www.nhu.edu.tw/~society/e-j/104/a40.htm
[30] https://bit.ly/41xYSrh
[31] https://www.parenting.com.tw/article/5049351
[32] https://www.facebook.com/JayWangFB/posts/188393636138799/
[33] https://se.ntpc.edu.tw/12basic/Main/Main1_2
[34] https://www.tainan.gov.tw/news_content.aspx?n=13371&s=3749018
[35] https://www.ceec.edu.tw/
[36] https://bit.ly/2W0fdUq
[37] https://udn.com/news/story/6885/5380520
[38] https://www.parenting.com.tw/article/5094739
[39] https://bit.ly/3bxIuSn

1. Compute the "top" as the average raw score of the highest 1% among test takers.
2. Divide the number by 15 to get the scale interval.
3. Create a table to map raw scores to the 0 to 15 point scale.

For example, if a subject has the "top" of 93.45 (max 100), then the scale interval is 93.45/15 = 6.23, rounded to two decimal points. Zero point is reserved for absolute zero marks on the exam. A student will get 1 point for a raw score between 0.01 and 6.23, and 2 points for a raw score between 6.24 and 12.46, and so on. If we observe from the top of the scale, a student can get 15 points for a raw score between 87.23 and 100, and 14 points for a raw score between 81.00 and 87.22. Note that the threshold (87.23) to achieve 15 points is 14 times of the scale interval (6.23).

Starting in 2019, students may choose four of the five subjects for the GSAT, and the maximum possible score (i.e., full marks) is reduced from 75 to 60.[40] Students in the science track often choose to drop the social sciences subject in the GSAT, and students in the humanities track often choose to skip the physical sciences subject. High school students are also required to maintain a portfolio each semester to document their learning outcomes, and the portfolio will be evaluated by the colleges they apply to.[41]

### 2.2.2 Advanced Subjects Test (AST)

## Unfinished below

The original AST existed from 2002 to 2021, and it was almost always held on July 1st, 2nd, and 3rd each year.[42] In the regular admission phase, the AST scores are directly used for college placements (Shy et al., 2021), without consideration of the student's achievements outside the AST scores.[43]

There are ten subjects in the original AST: Chinese, English, mathematics A, mathematics B, physics, chemistry, biology, civics & social studies, geography, and history.

Students may register for ?? subjects.
Depending on specific college requirements.

Don't go deep into commentaries or how the society reacted to the changes.
Elaborate on the new regular admission rules starting in 2022

In 2022, the original AST was replaced with the new AST with major changes.[44] For instance, Chinese and English are removed from the new AST, and students are required to use the GSAT scores for these two subjects.[45] The dates were also changed to mid-July.[46]

We do not have sufficient data for this part yet.

New AST: Wiki link: http://tinyurl.com/ycvetzsd
Each subject in the new AST is still scored from 0 to 100 points as before.
Mathematics A and mathematics B: to be determined, each year is different.

Continue writing below

[40] http://tinyurl.com/3r5zrzhs
[41] https://www.108epo.com/courses.php
[42] https://bit.ly/2J7YxoW
[43] With limited exceptions of GSAT score requirements. https://udn.com/news/story/121570/4773243
[44] https://www.ceec.edu.tw/unithome?psid=0J129606443348127072
[45] https://www.cw.com.tw/article/5119369
[46] https://www.cw.com.tw/article/5119369

# 3 Data Description

We retrieved data from the SENIORHIGH (high school)[47] discussion section on PTT,[48] the largest terminal-based bulletin board in Taiwan.[49] The data are stored in the file `ptt_SENIORHIGH_data.csv`.[50] The data from PTT are more representative than if we had collected on our own, because almost anyone could 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. The records indicate that each respondent had taken the college entrance exam in the year 2015 or earlier, so we can safely assume that neither of them had taken the new form of high school entrance exam starting in 2014.

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 our friends to report their scores anonymously, this approach can result in a large selection bias. Many of our friends graduated from the same high school and/or college, so we are likely to have similar entrance exam scores. That's why we turned to the public forum PTT to reduce selection bias in the data collection.

Data in the real world are messy, and data scientists spend lots of time cleaning (preprocessing) the data, i.e., preparing the data for analysis.[51] But data cleaning is a necessary step for better analysis results, and there are some visualization examples that demonstrate the importance of preprocessing the data (Chai, 2020). Our dataset `ptt_SENIORHIGH_data.csv` is relatively clean, but we still have to recode and flag some anomaly values.

## 3.1 Flagging Values in Data

The dataset `ptt_SENIORHIGH_data.csv` contains 197 rows, and 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) of the high school entrance exam in Taiwan, ranging from 0 to 99.
- **College_Score**: Each person's General Scholastic Ability Test (GSAT) score, ranging from 0 to 75.

There are 6 missing values in **HighSchool_PR** and 3 missing values in **College_Score**, so each of them is recorded as "-1" (an invalid numerical value for the scores). In the data entry process, invalid scores like **HighSchool_PR** 100 are also treated as missing.

In some cases, the reported scores can be inaccurate based on the respondent's description, so we read the comments and manually created two indicators for this issue. Note that **inaccurate scores are still valid**, so we keep them in the data analysis.

- **HS_Inacc**: A "1" means the reported **HighSchool_PR** is inaccurate.
- **CS_Inacc**: A "1" means the reported **College_Score** is inaccurate.

For **HS_Inacc** and **CS_Inacc**, we set the missing values to "0" because the two flags are binary indicators. Otherwise, the missing values would show as `NA`, which are difficult to process in the code.

```
data = read.csv("ptt_SENIORHIGH_data.csv")

names(data)[1] = "pttID" # system read in as "i..pttID", need to correct this

data$HS_Inacc[is.na(data$HS_Inacc)] = 0
data$CS_Inacc[is.na(data$CS_Inacc)] = 0
```

[47] https://www.ptt.cc/bbs/SENIORHIGH/M.1432729401.A.995.html
[48] If you have a PTT account, you can log into the website using a browser. https://term.ptt.cc/
[49] https://en.wikipedia.org/wiki/PTT_Bulletin_Board_System
[50] https://github.com/star1327p/Exam-Scores-PTT/blob/master/ptt_SENIORHIGH_data.csv
[51] https://bit.ly/303IWxY

One obvious reason of inaccurate scores is that a respondent may comment that his/her scores are approximate. Moreover, some people reported their **HighSchool_PR** from the mock exam, rather than from the actual high school entrance exam. In 2012 and 2013, the Ministry of Education in Taiwan allowed students to apply for high schools with their grades in middle school. During that time, if a student got admitted to a high school using this method, he/she would not need to take the high school entrance exam.[52]

For **College_Score**, the approximation clause still applies. In addition, there are two college entrance exams in each school year, and some students may do much better on the second exam than the first one. Then they were admitted to a more prestigious school than the first exam score had indicated, so this is also a form of inaccuracy.

We add two more indicators to signal invalid (impossible) values in **HighSchool_PR** and **College_Score**.

- **HS_Invalid**: A "1" means the reported **HighSchool_PR** is invalid, i.e., outside the range of 1-99.
- **CS_Invalid**: A "1" means the reported **College_Score** is invalid, i.e., outside the range of 1-75.

Again, we set **HS_Invalid** to "0" for a valid **HighSchool_PR**, and set **CS_Invalid** to "0" for a valid **College_Score**.

```
data$HS_Invalid = 0
data$HS_Invalid[(data$HighSchool_PR > 99)|(data$HighSchool_PR < 1)] = 1

data$CS_Invalid = 0
data$CS_Invalid[(data$College_Score > 75)|(data$College_Score < 1)] = 1
```

Finally, we save the clean version to a separate `.csv` file for future use.

```
write.csv(data, "ptt_SENIORHIGH_data_clean.csv")
```

## 3.2 Data Snapshot

We show the first 10 rows of data here, and we already see several anomalies that are flagged. For example, the 1st respondent `game275415` provided an approximate value to the **HighSchool_PR**, so we marked this value as inaccurate and a "1" in **HS_Inacc**. The 4th respondent `heejung` used mock exam scores for the **HighSchool_PR**, so this score is also marked as inaccurate. The 6th respondent `robinyu85` claimed to not having taken the high school entrance exam at all, so his/her missing **HighSchool_PR** is encoded to "-1" and flagged as invalid.

We also observed that **pttID** contains some information for potential inference, although we are not going to use it. For example, the 6th respondent `robinyu85` could be someone named Robin Yu, and the 8th respondent `godpatrick11` may have the English name Patrick. Nevertheless, this kind of information is simply a heuristic, so it is neither sufficient nor appropriate to include in the data analysis.

```
data[1:10,]
```

```
##           pttID HighSchool_PR College_Score HS_Inacc CS_Inacc HS_Invalid CS_Invalid
## 1    game275415            60            50        1        0          0          0
## 2      a2654133            60            52        0        0          0          0
## 3   cookie20125            99            72        0        0          0          0
## 4       heejung            92            54        1        0          0          0
## 5        shun01            87            51        0        0          0          0
## 6     robinyu85            -1            74        0        0          1          0
## 7    allengoose            69            48        0        0          0          0
## 8  godpatrick11            98            60        0        0          0          0
## 9     morgankhs            95            65        0        0          0          0
## 10      jazzard            88            65        0        0          0          0
```

[52] https://tsjh301.blogspot.com/2014/06/compulsory-education.html

### 3.3 Bivariate Validation

We need to clean up the data to prepare for the analysis. For example, we would like to investigate the relationship between **HighSchool_PR** and **College_Score**, so we need to ensure that each record consists of both valid scores. There are 191 records of valid **HighSchool_PR** numbers in the data, but if we consider only the ones with a valid **College_Score**, the number of available records drops to 188. Although which version we use does not matter much when we look at the univariate distribution, this will be problematic when we combine the univariate analysis with the bivariate analysis. Thus, we should use only the 188 records whose **College_Score** numbers are also valid.

```
# High school entrance exam scores: Keep only valid numbers
uni_HS_score = data$HighSchool_PR[which(data$HS_Invalid == 0)]
length(uni_HS_score)
```

```
## [1] 191
```

The same requirement also applies to **College_Score**. There are 194 records of valid **College_Score** numbers in the data, but only 188 of them also have corresponding valid **HighSchool_PR** numbers.

```
# College entrance exam scores: Keep only valid numbers
uni_college_score = data$College_Score[which(data$CS_Invalid == 0)]
length(uni_college_score)
```

```
## [1] 194
```

We do not have enough information to impute the few missing values in our dataset, so we decided to exclude them from our analysis. There are nine records with invalid or missing values, and we can locate their indices in the data.

```
missing_rows = which(data$HS_Invalid == 1 | data$CS_Invalid == 1)
missing_rows
```

```
## [1]   6  19  71  85  88  96 132 183 195
```

**Remark**: With larger and/or more complex datasets, researchers often apply record linkage methods (Dusetzina et al., 2014) to impute missing or erroneous values in one dataset from another. This is outside the scope of this manuscript.

Now we build the `R` dataframe `data_corr` (short for "data correlation") for the bivariate analysis. This dataframe excludes any record with at least one missing value in **HighSchool_PR** or **College_Score**.

```
# Remove missing data
data_corr = data[-missing_rows,]
```

Again, we show the first 10 rows of `data_corr`, and all values in `HighSchool_PR` and `College_Score` are valid. The indicators `HS_Invalid` and `CS_Invalid` should be zero throughout this version of data.

```
data_corr[1:10,]
```

```
##           pttID HighSchool_PR College_Score HS_Inacc CS_Inacc HS_Invalid CS_Invalid
## 1    game275415            60            50        1        0          0          0
## 2      a2654133            60            52        0        0          0          0
## 3   cookie20125            99            72        0        0          0          0
## 4       heejung            92            54        1        0          0          0
## 5        shun01            87            51        0        0          0          0
## 7    allengoose            69            48        0        0          0          0
## 8  godpatrick11            98            60        0        0          0          0
## 9     morgankhs            95            65        0        0          0          0
## 10      jazzard            88            65        0        0          0          0
## 11     ksacball            99            70        0        0          0          0
```

The function `dim` retrieves the dimensions of the `R` dataframe `data_corr`. The output reveals that the dataset contains 188 rows and 7 columns, where the 188 rows refer to the 188 records with a valid score in both `HighSchool_PR` and `College_Score`.

```
dim(data_corr)
```

```
## [1] 188   7
```

Alternatively, we can also use the function `length` to find the number of elements in an array. This shows both columns `HighSchool_PR` and `College_Score` contain 188 elements each. Note that `length` does not distinguish between missing and non-missing values. A missing element, often denoted as `NA`, is still counted as one element.

```
length(data_corr$HighSchool_PR)
```

```
## [1] 188
```

```
length(data_corr$College_Score)
```

```
## [1] 188
```

**Remark**: Readers new to `R` programming may wonder where we can find the type of each object, in order to determine how to proceed next. The function `class` returns the type of the input object, and the output confirms that `data_corr` is an `R` dataframe. Then we can leverage the functions available for this data structure.

```
class(data_corr)
```

```
## [1] "data.frame"
```

# 4 Exploratory Data Analysis

Before we perform any statistical modeling, we need to observe the data and this step is called exploratory data analysis. The exploratory phase allows us to identify patterns, detect anomalies, and verify assumptions in the data.[53] This is an important but often overlooked step in data analysis, and failure to explore the data can lead to inefficiencies such as accurate models on the wrong data.[54]

Therefore, we start with examining the trends of the two main variables, **HighSchool_PR** and **College_Score**. For each variable, we observe the values and summarize them in a histogram as part of the univariate analysis. Then we investigate the relationship between the two variables, i.e., bivariate exploration.

## 4.1 High School Entrance Exam Scores (Percentile Rank)

We show the descriptive statistics of **HighSchool_PR** in our data, i.e., the percentile rank of high school entrance exam scores. The invalid or missing values are removed beforehand. The `summary` function returns the minimum, 1st quartile, median, mean, 3rd quartile, and maximum from the input data.

- **Minimum**: The smallest value in the data.
- **1st Quartile**: The value where 25% of the data is below this point; that is, the middle value between the minimum and the median.
- **Median**: The value where 50% of the data is below this point, and 50% of the data is above this point.
- **Mean**: The arithmetic mean, which is calculated as the sum of the data divided by the number of the points.
- **3rd Quartile**: The value where 75% of the data is below this point; that is, the middle value between the median and the maximum.
- **Maximum**: The largest value in the data.

[53] https://blog.eduonix.com/bigdata-and-hadoop/importance-exploratory-data-analysis-ml-modelling
[54] https://towardsdatascience.com/exploratory-data-analysis-topic-that-is-neglected-in-data-science-projects-9962ae078a56

```r
summary(uni_HS_score)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   51.00   85.00   92.00   89.82   97.00   99.00
```

**Remark**: Note that we used the full univariate set of **HighSchool_PR** available, rather than taking only those with a valid **College_Score**. The reason is that we wanted to keep as many records as possible. We can do a quick sensitivity analysis here, and the summary statistics are very similar after we exclude the few records without a valid **College_Score**.

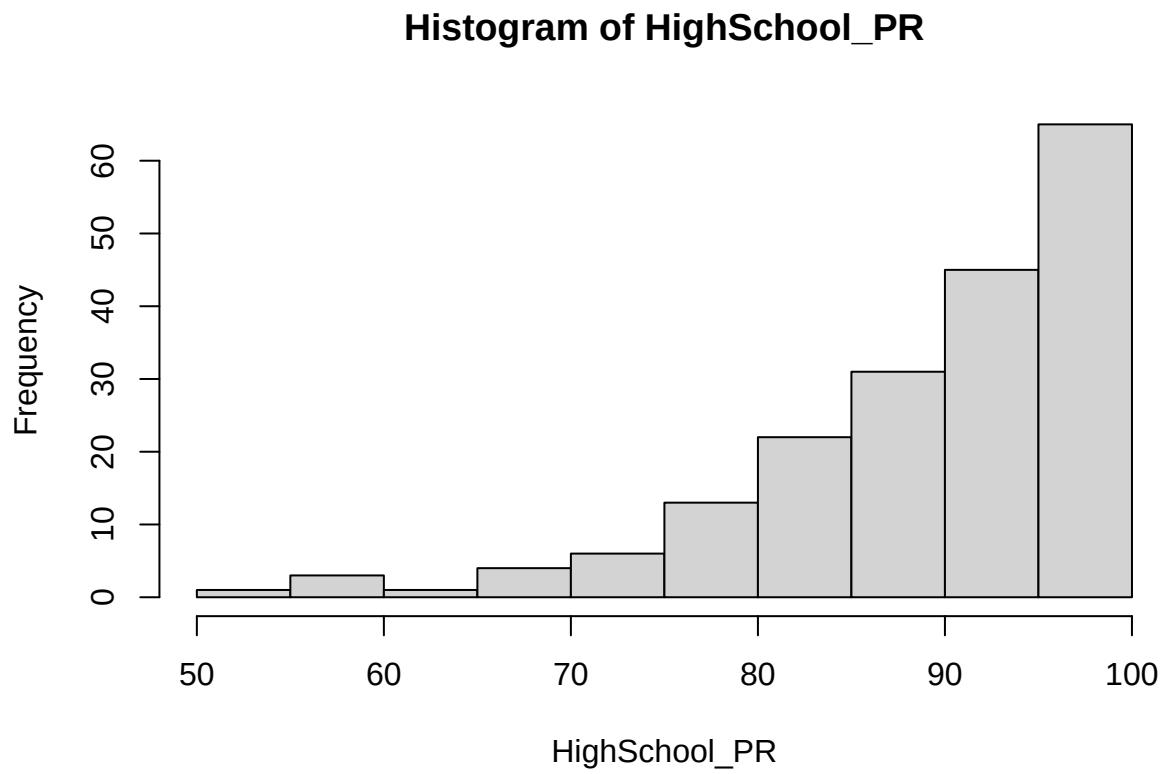
```r
summary(data_corr$HighSchool_PR)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    51.0    85.0    92.0    89.7    97.0    99.0
```

In the univariate analysis, the 1st quartile of **HighSchool_PR** is 85, which means only 25% of the respondents have a percentile rank (PR) below 85. Hence 75% of the respondents have a percentile rank (PR) at least 85, i.e., in the top 15% of the high school entrance exam. The median (92) is higher than the mean (89-90), indicating that the distribution is left-skewed. Half of the respondents scored **HighSchool_PR** 92 or better, and we have few samples of the lower end of the high school entrance exam. In fact, the lowest score we obtained in the data is already 51 – still slightly above the national median score (50). The maximum is capped at 99.

The histogram shows a left-skewed distribution, and we can see a long tail to the left. Since our data are self-reported by respondents, the ones with higher scores are more likely to report, resulting in survey non-response bias. It is also possible that many students with extremely low **HighSchool_PR** chose not to attend college, so they would not have a **College_Score** to respond to the survey.

```r
hist(uni_HS_score, main = "Histogram of HighSchool_PR", xlab="HighSchool_PR")
```




If we were to take a sufficiently large random sample from the full database of **HighSchool_PR** in Taiwan, the minimum should be 1 and the maximum would still be at 99. However, the **HighSchool_PR** is the national percentile rank itself, so the 1st quartile should be around 25, median and mean both around 50, and the 3rd quartile around 75.

As a baseline, we also simulate this random sample by generating data from a **discrete** uniform distribution between 1 and 99. We manually take a random sample of the same size as the univariate **HighSchool_PR** from the integers 1, 2, . . . , 99, and we sample with replacement to allow duplicate values. Note that the function `runif`[55] in the `stats` package generates samples from the **continuous** uniform distribution, and that's why we did not use `runif` here.

```
possible_values = c(1:99)

set.seed(67)
# sample with replacement
random_sample = sample(possible_values, size=length(uni_HS_score), replace=TRUE)
print(random_sample)
```

```
##   [1] 43 84 97 78 71 33 94 45 52 45 32  1 68 80 24  5 96 27 62  3 24 48 63 85 52
##  [26] 99 79 30 73 63 59 19 10 70 67 77 76 59 80 11  4 43 12 33 22 45 12 68 23 98
##  [51] 13  1 96 28 53 74 19 11 42 77 75 31 94 15 31 75 76 63 37 77 52 86 64 15 32
##  [76] 97  4 18 11 16 36 70 18 91 23 39 70 58 71 90  4 32 91 19 83 16 67 88 71 16
## [101] 10  7 65 42 56 19 79  7 26 74  4 32 83 98 77 61 49 66 46  6 17 48 84 41 65
## [126] 64 38 45 79 18 55 23 37 69  9 63 41 73 35 55  2 38 23 58 16 91 94 94 16 62
## [151]  3 57 49 36 84 94 45 62 53 29 81 73 87 22 99 36 29 82  6 84  7 65 18 33 68
## [176] 35 84 67  3 57 24 13 77 14 49 25 83 41  8 68  6
```

[55]The function name `runif` is pronounced as `r-unif`, not `run-if`.

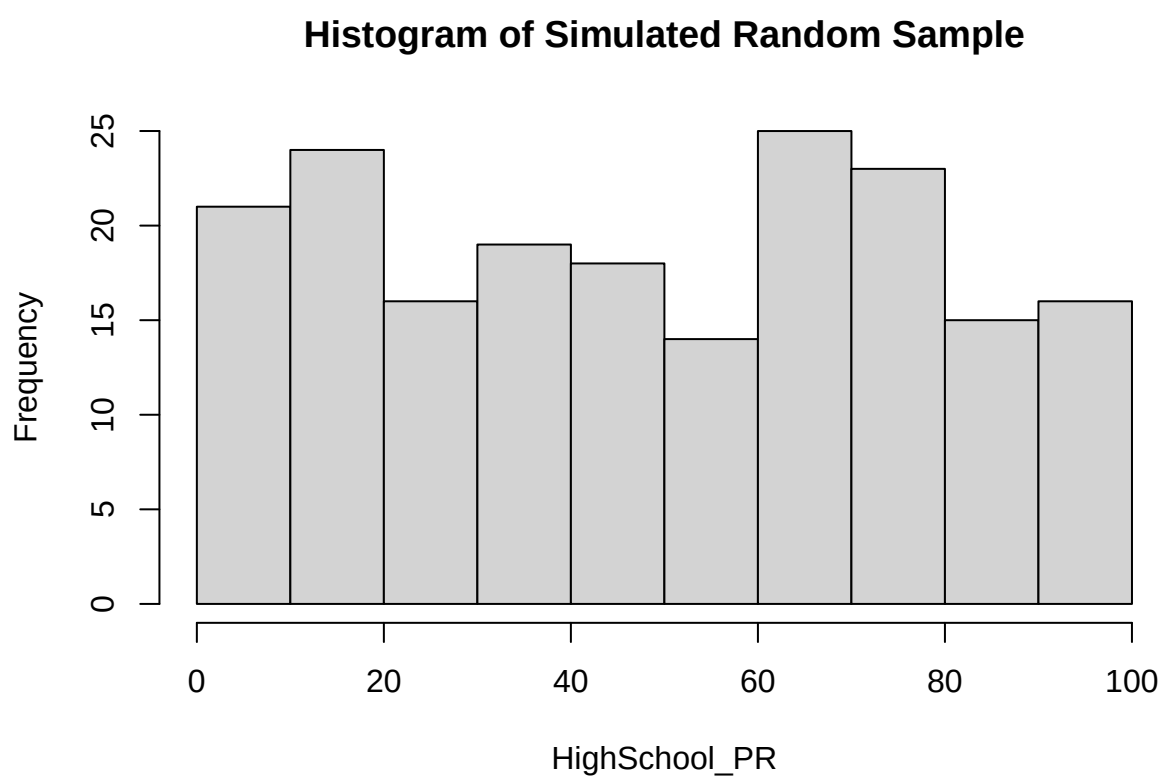
As we can see, the simulated random sample has median and mean close to 50. The 1st quartile is near 25, and the 3rd quartile is near 75. The simulated sample is to demonstrate a nationally representative sample of **HighSchool_PR**. Nevertheless, this would not be appropriate for the evaluation of the relationship between **HighSchool_PR** and **College_Score**, because students with extremely low **HighSchool_PR** may decide not to attend college at all.

```
summary(random_sample)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00   23.00   49.00   48.67   73.50   99.00
```

The histogram of the simulated random sample is flat, as what we expect from a discrete uniform distribution.

```
hist(random_sample, main = "Histogram of Simulated Random Sample", xlab="HighSchool_PR")
```




## 4.2 College Entrance Exam Scores

Similarly, we also show the descriptive statistics of **College_Score**, i.e., the college entrance exam scores between 1 and 75. The distribution is also left-skewed, but less extreme than **HighSchool_PR**. The median of **College_Score** is 64.5, indicating that 50% of the respondents have **College_Score** 65 or higher. The 3rd quartile is already at 69, so the top score range 69-75 accounts for 25% of the respondents. This is much higher than the national average.

```
summary(uni_college_score)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    34.0    58.0    64.5    62.7    69.0    75.0
```

As a sensitivity analysis check, we also examine the `summary` of the **College_Score** datapoints only when

they have a corresponding valid **HighSchool_PR**. The distribution is almost the same.

```
summary(data_corr$College_Score)
```

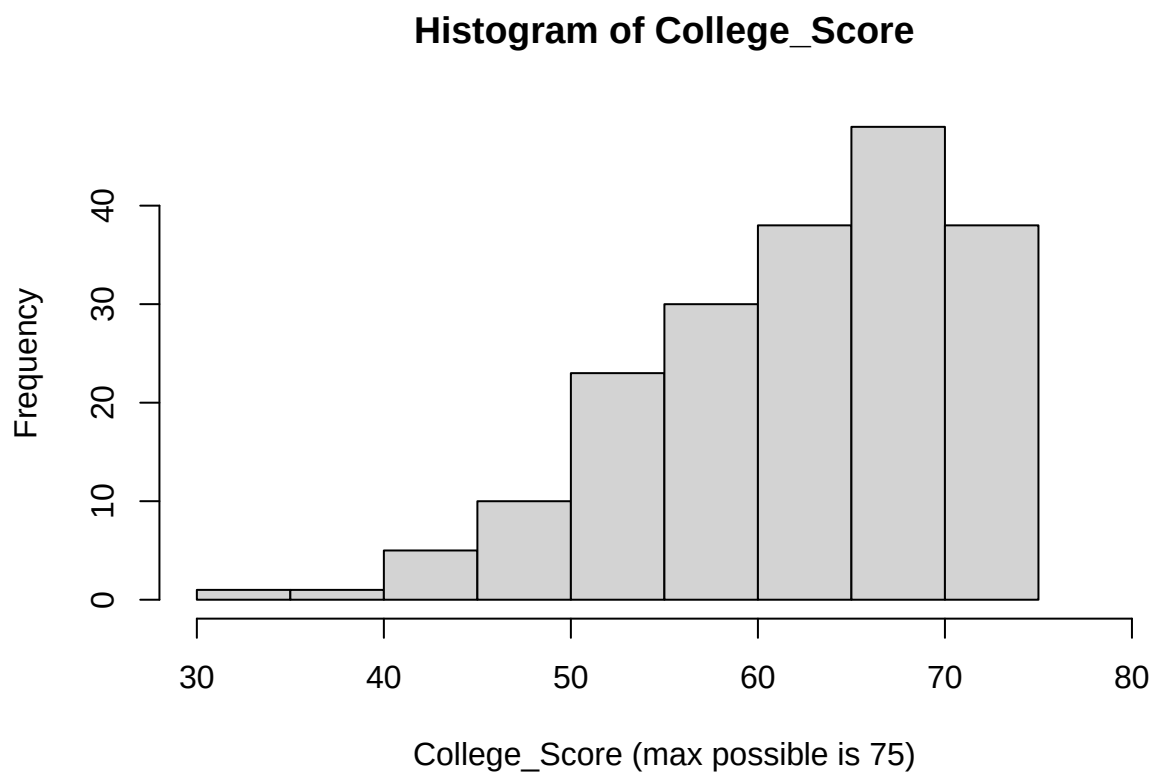
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   34.00   58.00   64.00   62.49   69.00   75.00
```

According to the reference score table[56] from Wikipedia, the 88th percentile of the college entrance score fluctuates around 60 in Years 2004-2010, and 62-65 in Years 2011-2018. Since the median of **College_Score** is 64.5, we can infer that at least 50% of the respondents scored in the top 12% of the college entrance exam. On the other hand, the reference score table also shows that the 75th percentile of the college entrance score is between 53 and 58 in Years 2004-2018. The PTT data's 1st quantile is already at 58, so we can also infer that at least 75% of the respondents scored in the top 25% of the college entrance exam.

Since PTT contains forums for several prestigious universities in Taiwan, it is no surprise that many users attended these colleges because they scored well on the college entrance exam. Nevertheless, PTT does not limit registration to students of these colleges, so the population of PTT is relatively diverse but still not very representative of the whole student population.

The histogram for **College_Score** also shows a left-skewed distribution, but not as extreme as **HighSchool_PR**. Note that right-skewed distributions are more common in real life, such as employee salaries and movie ticket sales.[57]

```
hist(uni_college_score, main="Histogram of College_Score",
     xlab="College_Score (max possible is 75)",xlim=c(30,80))
```




Unlike the case of **HighSchool_PR**, we do not think it is appropriate to simulate a random sample for **Col-**

[56] https://bit.ly/3bAYOvO
[57] https://www.statology.org/positively-skewed-distribution-examples/

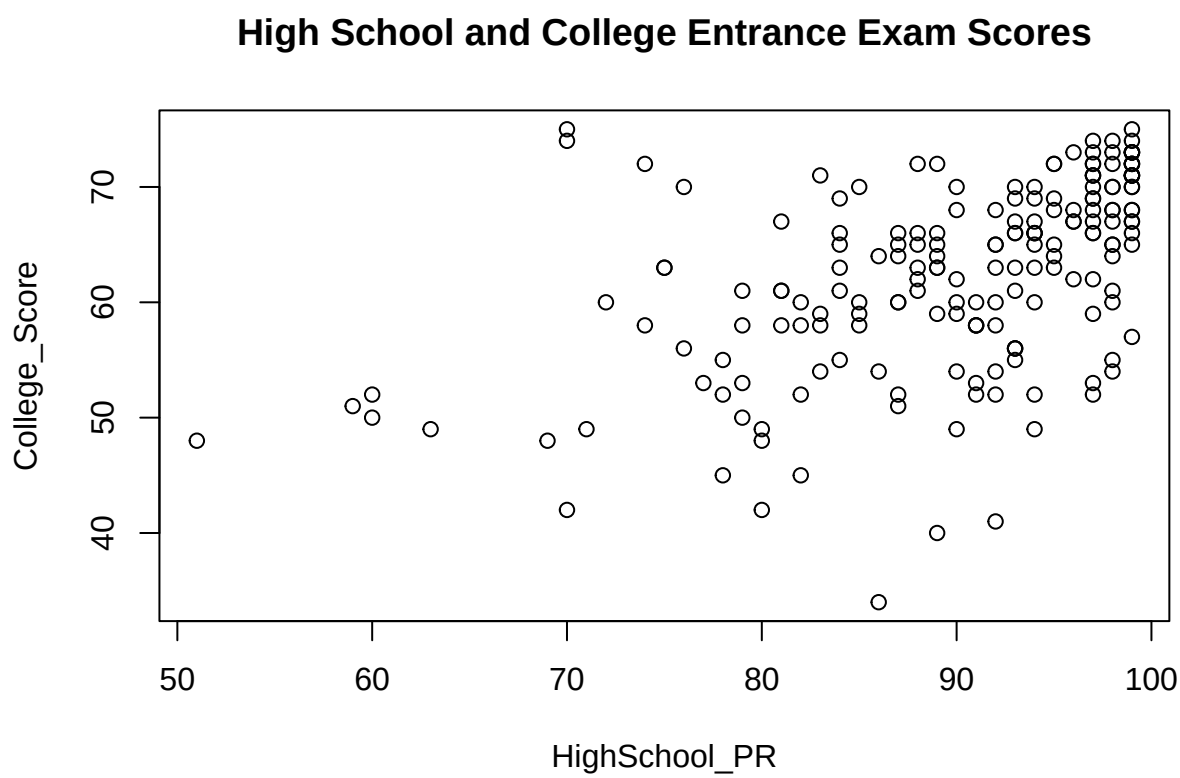
**lege_Score** across years in general. First, **College_Score** are not percentile ranks like **HighSchool_PR**, and there is no statistical nature of how the **College_Score** distribution should look like. Moreover, the exact distribution of **College_Score** differs greatly due to the varying difficulty levels each year. For example, the 88th percentile of **College_Score** was 65 in 2014 and 59 in 2007.[58] The 6-point difference in the score is large enough for an applicant to get admitted to a lower or higher tier of universities.

In Section 2.2, we mentioned that the Taiwan government published detailed statistics of the **College_Score** each year, including the five indicators (88th, 75th, 50th, 25th, 12th percentile) of each subject and the total score. However, since our data sample contains **College_Score** across several years, it is not meaningful to generate a simulated distribution here.

## 4.3 Bivariate Exploration

Next, we create a bivariate scatterplot of **HighSchool_PR** and **College_Score** to examine the relationship between the two scores. Obviously, a respondent needs a valid score in both to be included in the bivariate scatterplot. Just like what we observed in the univariate plots, both variables are largely concentrated towards the maximum possible scores. The bivariate scatterplot shows a funnel shape – respondents with lower **HighSchool_PR** have higher variance in **College_Score**.

```
plot(data_corr$HighSchool_PR, data_corr$College_Score,
     main = "High School and College Entrance Exam Scores",
     xlab="HighSchool_PR",
     ylab="College_Score")
```




We use the correlation coefficient to measure the strength of the linear relationship between **HighSchool_PR** and **College_Score**. The computed value is approximately 0.507, showing a medium strength of positive

[58] https://bit.ly/2W0fdUq

association between the two variables. We can interpret that a better score in the high school entrance exam is associated with a better college entrance exam score, but the relationship is not as strong after **HighSchool_PR** reaches 80. (Note that correlation does not imply causality!)

```
cor(data_corr$HighSchool_PR, data_corr$College_Score)
```

```
## [1] 0.5074473
```

To find the correlation coefficient between the random variables $X, Y$, we start with the covariance $\text{Cov}(X, Y)$ in the equation below. $E[X]$ denotes the expectation value of $X$, a.k.a. the mean of $X$.

$$\text{Cov}(X, Y) = E[(X - E[X])(Y - E[Y])] = E[XY] - E[X]E[Y].$$

Then we also need to compute the standard deviation $\sigma_X$:

$$\sigma_X = \sqrt{E[(X - E[X])^2]} = \sqrt{E[X^2] - (E[X])^2}.$$

Similarly, the standard deviation $\sigma_Y$ is:

$$\sigma_Y = \sqrt{E[(Y - E[Y])^2]} = \sqrt{E[Y^2] - (E[Y])^2}.$$

Finally, we can calculate the correlation coefficient as:

$$\rho_{X,Y} = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}.$$

The correlation coefficient $\rho$ is always between -1 and +1. The stronger $|\rho_{X,Y}|$ is, the stronger is the linear relationship. A positive value means the two variables $X, Y$ are positively associated with each other. In other words, when $X$ increases, $Y$ is also expected to increase. A negative value means the two variables $X, Y$ are negatively associated with each other. In this case, when $X$ increases, $Y$ is expected to decrease. The correlation coefficient may be zero, but this simply means that $X$ and $Y$ do not have a linear relationship with each other. There may exist other patterns between the two variables.

The strength of linear relationship can be interpreted as below:[59]

- $|\rho| = 1$: Perfect linear relationship
- $0.6 < |\rho| < 1$: Strong linear relationship
- $0.4 < |\rho| < 0.6$: Moderate linear relationship
- $0.2 < |\rho| < 0.4$: Low linear relationship
- $0 < |\rho| < 0.2$: Little-to-no linear relationship
- $|\rho| = 0$: No linear relationship

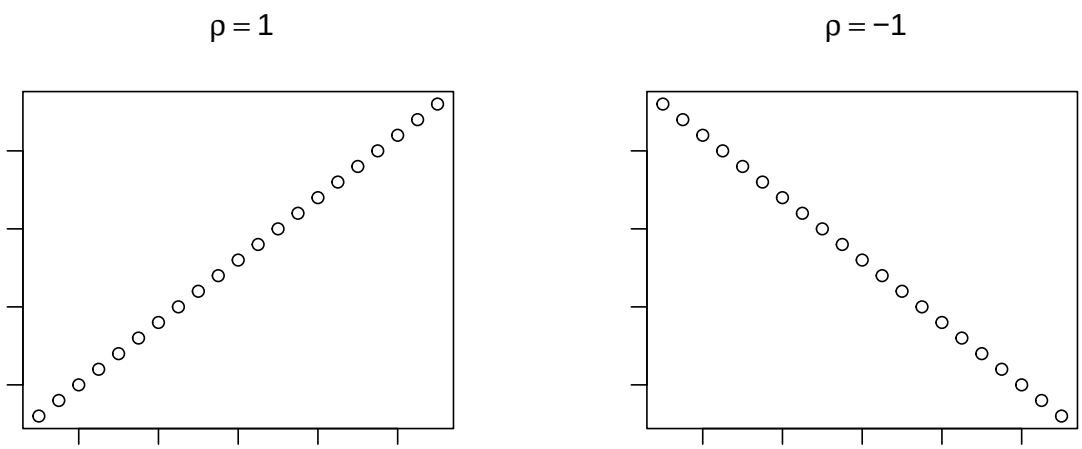
### 4.3.1 Examples of the Correlation Coefficient

Here we visualize some bivariate patterns for different values of the correlation coefficient $\rho$, with hypothetical data and examples. The graph ideas came from *OpenIntro Statistics* (Diez et al., 2019). We hide the code that generated these graphs to avoid obscuring the topic, and the interested readers can find the code in the raw `.Rmd` files. We will discuss more about linear regression in Chapter 5.

The first row of graphs show what $|\rho| = 1$ looks like – a straight line, i.e., a perfect linear relationship. In mathematical terms, we can write $Y = \alpha + \beta X$, where the sign of $\beta$ controls the direction of the pattern.

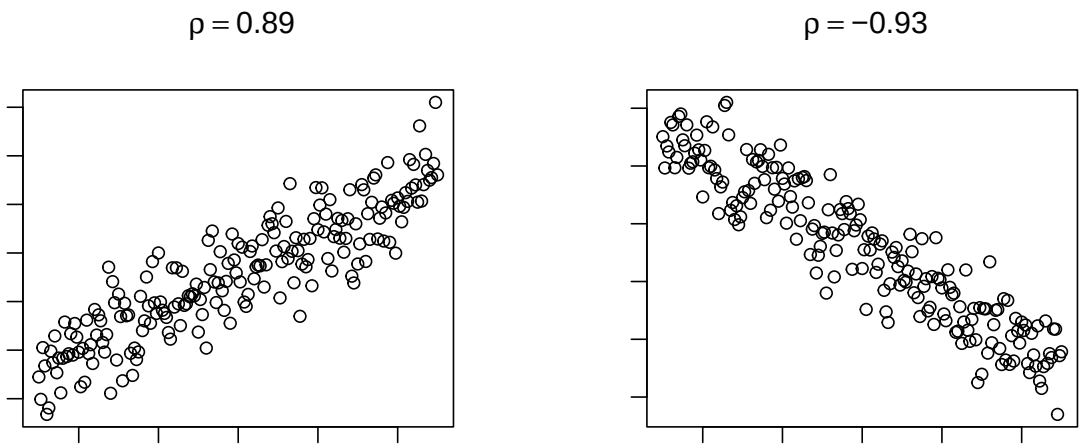
[59] https://bit.ly/3AUI8PA

- For $\rho = 1$, the two variables $X, Y$ have a perfect positive linear relationship ($\beta > 0$). One example is people's height in cm and height in inches, where the two variables are a positive linear combination of each other.
- For $\rho = -1$, the two variables $X, Y$ have a perfect negative linear relationship ($\beta < 0$). One example is you eat from a box of 12 cookies. The cookies you ate and the cookies remaining should add up to 12, so they have a perfect negative linear relationship.



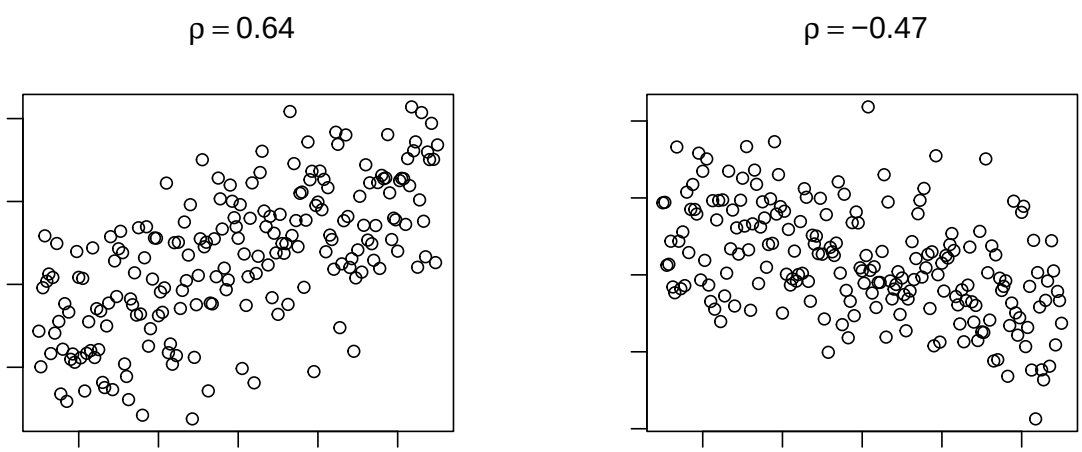

In the second row, the two variables in each plot have a good but imperfect linear relationship. The datapoints seem to be on a straight line with some fluctuation. If we start with the linear equation, we can add a small random noise with mean zero to the data generative process. Hence the equation becomes $Y = \alpha + \beta X + \epsilon$, where $\epsilon \sim N(0, \sigma^2)$. We assume the random noise $\epsilon$ follows a normal distribution, although this is not always necessary.

- The left graph illustrates $\rho = 0.89$. The plot shows a clear and positive relationship with some noise. One example is the height and weight of children of various ages. Taller children usually weigh more, but this is not always the case. (Childhood obesity is a serious health problem.)
- The right graph illustrates $\rho = -0.93$. The plot shows a clear and negative relationship with some noise. One example is the hours for work and the hours for hobbies. Generally, the more time you spend on work, the less time you spend on hobbies. Although everyone has 24 hours in a day, the time outside work does not always mean you are spending on hobbies for enjoyment.




The third row contains graphs that visualize a medium level of linear correlation, where $0.4 < |\rho| < 0.6$. The linear trend is still visible, but with more fluctuation than in the high-correlation graphs.

- The left figure shows $\rho = 0.64$, as a positive medium correlation of the two variables. One example is students' midterm exam scores and final exam scores. Students who did well on the midterm often continue to do well on the final, but this is not guaranteed. Exam scores often have more variability than people's height and weight.
- The right figure shows $\rho = -0.47$, as a negative medium correlation of the two variables. One example is adults' muscle mass and their age. Older adults tend to have less muscle than younger ones, but it is well-known that strength training helps preserve some muscle mass.[60]

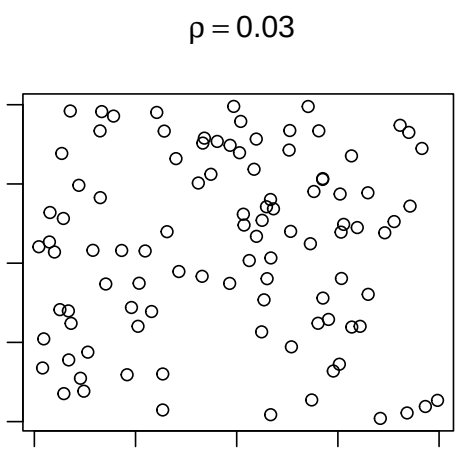


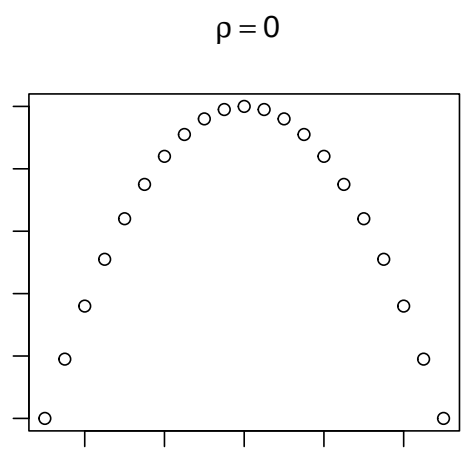

The final row shows what little-to-no linear correlation looks like. We need to remember that zero correlation

[60]We are not medical professionals, so we make the caveat that the linear relationship may not be accurate.

does not mean the two variables are independent. They may have a non-linear relationship, which cannot be detected by the correlation coefficient.

- The left image shows a bivariate scatter plot with $\rho = 0.03$, i.e., very little correlation. For a small $\rho$ in absolute value, the two variables barely have any linear relationship. Although the bivariate scatter plot looks random, the correlation coefficient may not be exactly zero. One example is the daily coffee consumption and intelligence (IQ scores). How much coffee an individual drinks is completely unrelated to his/her intelligence level, so we cannot use one variable to obtain information of the other.[61]
- The right image has $\rho = 0$ (absolute zero), i.e., no linear relationship at all. However, the graph shows a parabola curve, so the two variables have some non-linear relationship. If you throw a ball to another person, the ball's trajectory would look like a parabola.[62] The x-axis is horizontal and parallel to the ground, while the y-axis is vertical and perpendicular to the x-axis.




# 5 Linear Regression

Linear regression is a commonly-used statistical model to evaluate the relationship between two continuous variables, and even major companies like IBM endorse linear regression for better decision-making in business.[63] Therefore, we are going to decide whether we should run a linear regression to predict **College_Score** from **HighSchool_PR**. If yes, we would implement the model and check the residuals. If no, we need to explore other options in analyzing the data.

A linear regression model can be written in mathematical terms:

$$Y = \alpha + \beta X + \epsilon$$

$Y$ is the response variable, i.e., what we would like to predict. $X$ is the explanatory variable, i.e., the data used to make the predictions. $\alpha$ is the intercept, and it stands for the estimate $\hat{Y}$ when $X = 0$ (if applicable). $\beta$ is the coefficient, and when $X$ increases by one unit, we can expect $Y$ to increase by $\beta$ units. Last but not least, $\epsilon$ is the error term, which is normally distributed with mean zero.

*OpenIntro Statistics* (Diez et al., 2019) states that four requirements need to be met in a linear regression:

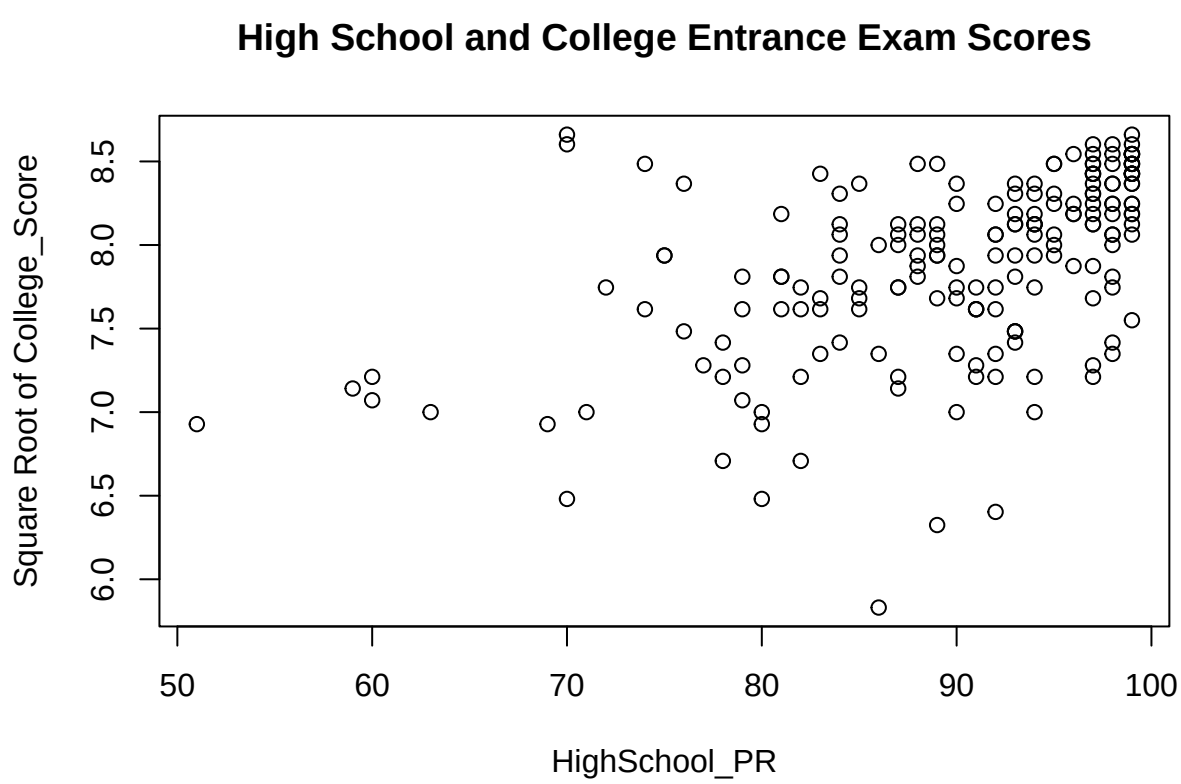
[61] https://www.statology.org/no-correlation-examples/
[62] https://www.wondriumdaily.com/mathematics-of-falling-the-parabolic-movement/
[63] https://www.ibm.com/topics/linear-regression

1. **Linearity**: The data has a linear trend, not a curve.
2. **Nearly normal residuals**: The residuals should be nearly normal, and we need to beware of outliers and influential points.
3. **Constant variability**: The variability of $Y$ is constant and does not depend on the value of $X$.
4. **Independent observations**: Each observation (datapoint) is independent of the others.

## 5.1 Should we run a linear regression?

It is inappropriate to perform linear regression directly, because the data do not meet the constant variability assumption. In the bivariate exploratory plot, we can see that the variability of **College_Score** ($Y$) increases as **HighSchool_PR** ($X$) increases. One possible remedy is apply the square root transformation to **College_Score**, in order to reduce the variability. But the scatterplot below shows little to no improvement in variability, and the correlation coefficient even drops from 0.507 to 0.504. Hence we determine that it is not a good idea to run a linear regression model on the whole dataset.

```
# data version: already removed missing data
# data_corr = data[-missing_rows,]

plot(data_corr$HighSchool_PR, sqrt(data_corr$College_Score),
     main = "High School and College Entrance Exam Scores",
     xlab="HighSchool_PR",
     ylab="Square Root of College_Score")
```



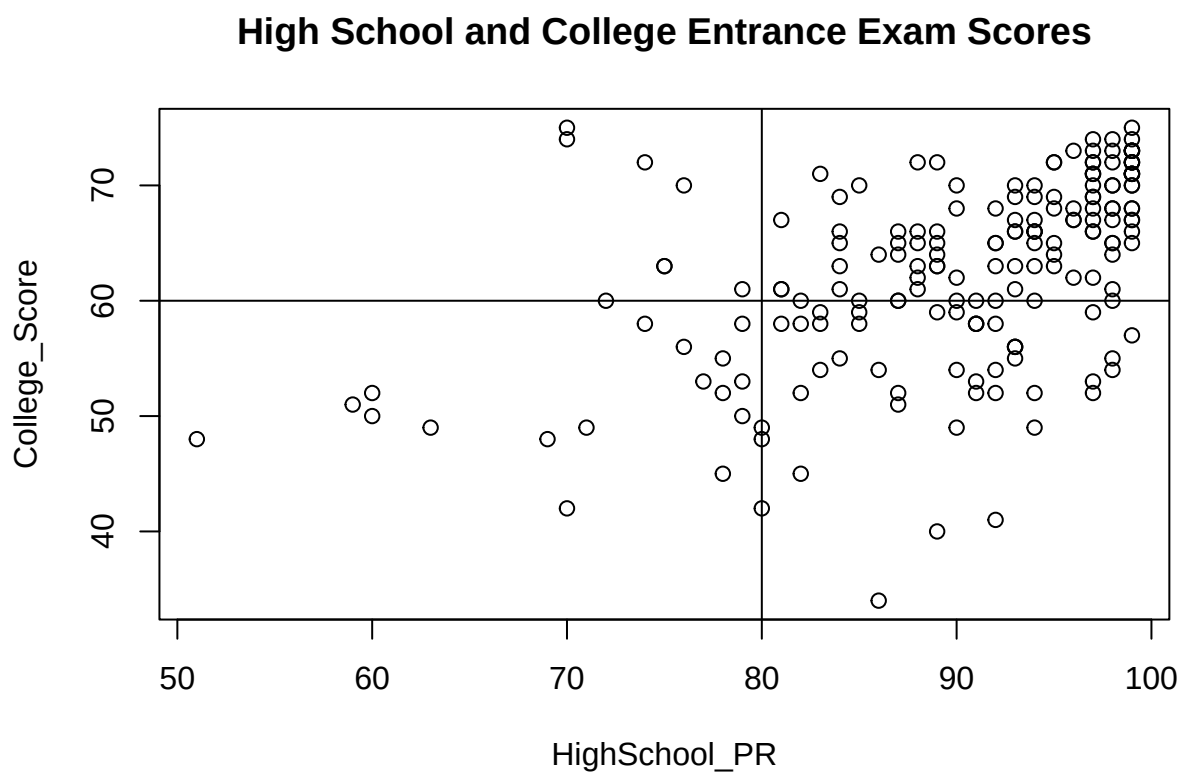

```
cor(data_corr$HighSchool_PR, sqrt(data_corr$College_Score))
```

```
## [1] 0.5048132
```

## 5.2 Segmenting the Data

Instead, we should segment the data and further examine the top scorers in the dataset, i.e., those with **HighSchool_PR** 80 or higher. Most of these respondents have **College_Score** of 60 or higher, but the range of **College_Score** is wide. Here, we add horizontal and vertical lines to clarify the graph.

```
plot(data_corr$HighSchool_PR, data_corr$College_Score,
     main = "High School and College Entrance Exam Scores",
     xlab="HighSchool_PR",
     ylab="College_Score")

abline(h=60,v=80)
```




We can also create a contingency table (a.k.a. cross tabulation) for the two indicators **HighSchool_80up** and **College_60up**, which displays the bivariate frequency distribution in terms of counts.

- **HighSchool_80up**: Indicator of whether **HighSchool_PR** is 80 or higher
- **College_60up**: Indicator of whether **College_Score** is 60 or higher

```
data_corr$HS_80up = data_corr$HighSchool_PR >= 80
data_corr$CS_60up = data_corr$College_Score >=60

contingency = table(data_corr$HS_80up, data_corr$CS_60up,
                    dnn=c("HighSchool_80up","College_60up"))

contingency
```

```
##                College_60up
## HighSchool_80up FALSE TRUE
```

```
##            FALSE    17    8
##            TRUE     43  120
```

To make the table easier to read, we revert the order of `FALSE` and `TRUE` in the contingency table by calling the indices in reverse order.

```
contingency = contingency[2:1, 2:1]
contingency
```

```
##                College_60up
## HighSchool_80up TRUE FALSE
##            TRUE   120    43
##            FALSE    8    17
```

Below is the percentage version of the contingency table, and we can see that more than 63.5% of the respondents have both **HighSchool_PR** $\geq 80$ and **College_Score** $\geq 60$. This is also evidence that the PTT users tend to come from the population who scored well on the high school and college entrance exams.

```
prop.table(contingency)
```

```
##                College_60up
## HighSchool_80up       TRUE      FALSE
##            TRUE  0.63829787 0.22872340
##            FALSE 0.04255319 0.09042553
```

We can also round the percentage version to four decimal places in the ratio, so we will have two decimal places after the integer percentage. For example, 0.4528 becomes 45.28%.

```
round(prop.table(contingency),4)
```

```
##                College_60up
## HighSchool_80up   TRUE  FALSE
##            TRUE  0.6383 0.2287
##            FALSE 0.0426 0.0904
```

## 5.3 Conditional Probability

Using conditional probability, we can answer this question from the data: If a person scores at least 80 on the high school entrance score percentile rank (PR), how likely is he/she going to obtain a score at least 60 on the college entrance exam?

In mathematical terms, this is equivalent to finding $P(\text{College_60up is true} \mid \text{HighSchool_80up is true})$. Recall the conditional probability formula and the Bayes theorem:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B|A)P(A)}{P(B)}$$

In this data, we have

- $P(\text{HighSchool_80up is true})$ = # of respondents with **HighSchool_PR** $\geq 80$ / all respondents.
- $P(\text{College_60up is true})$ = # of respondents with **College_Score** $\geq 60$ / all respondents.
- $P(\text{HighSchool_80up is true} \cap \text{College_60up is true})$
  = # of respondents with **HighSchool_PR** $\geq 80$ and **College_Score** $\geq 60$ / all respondents.

Plugging the numbers into the equation, we get

$$
\begin{aligned}
&P(\text{College_60up is true} \mid \text{HighSchool_80up is true}) \\
&= \frac{P(\text{HighSchool_80up is true} \cap \text{College_60up is true})}{P(\text{HighSchool_80up is true})} \\
&= \frac{\#\text{ of respondents with HighSchool_PR} \geq 80 \text{ and College_Score} \geq 60}{\#\text{ of respondents with HighSchool_PR} \geq 80} \\
&= \frac{120}{43+120} \approx 0.7362.
\end{aligned}
$$

According to this data from PTT, there is a 73.62% chance for a person to score at least 60 on the college entrance exam, given that he/she scored at least 80 on the high school entrance score percentile rank (PR). Note that we use number of respondents rather than percentage to avoid rounding errors.

In comparison, if we do not know anything about the person's high school entrance score percentile rank (PR), we have a probability of 63.82% in observing the person scoring at least 60 on the college entrance exam. There is an increase of 9.80% in probability after we learn information about his/her high school entrance exam score.

$$
\begin{aligned}
P(\text{College_60up is true}) &= \#\text{ of respondents with College_Score} \geq 60 \text{ / all respondents} \\
&= \frac{120}{188} \approx 0.6382.
\end{aligned}
$$

```r
nrow(data_corr) # number of all respondents without missing data
```

```
## [1] 188
```

**Remark**: Conditional probability is the foundation of Bayesian statistics, which updates the existing probabilities given the new data. For the interested readers, we recommend the book *An Introduction to Bayesian Thinking* (Clyde et al., 2018) as a start. This book was written as a companion for the Bayesian Statistics course on Coursera from Duke University.[64] Knowledge of calculus is helpful but not an absolute prerequisite.

# 6 Top Scorers: A Closer Look

We would like to further examine the top scorers, given the high number of records in the data observed in Section 5.2. In Taiwan's education system, the top tier of high schools and colleges can be further segmented. The top of the range can be very different than the middle or bottom. For example, National Taiwan University[65] is the most prestigious university in Taiwan, but Chang Gung University[66] is also excellent in medical education. However, many people in the US have heard of the former but not the latter. This does not mean you should always choose the most prestigious college; many other factors should also be considered, such as tuition and majors offered at that college.

We consider the following subcategories in the entrance exam scores:

- **HighSchool_PR** ranges: 80-89, 90-94, 95-99
- **College_Score** ranges: 60-64, 65-69, 70-75

Let's investigate the univariate distributions of **HighSchool_PR** and **College_Score** respectively, then we move on to explore the bivariate relationship between the two sets of scores. It is good practice to start with one variable at a time, even when we are mainly interested in the relationship between the two variables.

[64] https://www.coursera.org/learn/bayesian
[65] https://www.ntu.edu.tw/english/
[66] https://www.cgu.edu.tw/?Lang=en

## 6.1 High School Entrance Exam Scores (Percentile Rank) at least 80

We use the `R` function `table` to show the frequency of each **HighSchool_PR** value that is at least 80, and we have 163 values in total. In the table below, the first row is the PR (percentile rank), and the second row is the counts. Although we truncated the **HighSchool_PR** to 80 and above, the distribution is still left-skewed. The **HighSchool_PR** 99 has the highest counts, followed by **HighSchool_PR** 97 and **HighSchool_PR** 98.

```
HS_PR_seg = data_corr$HighSchool_PR[which(data_corr$HS_80up == TRUE)]
length(HS_PR_seg)
```

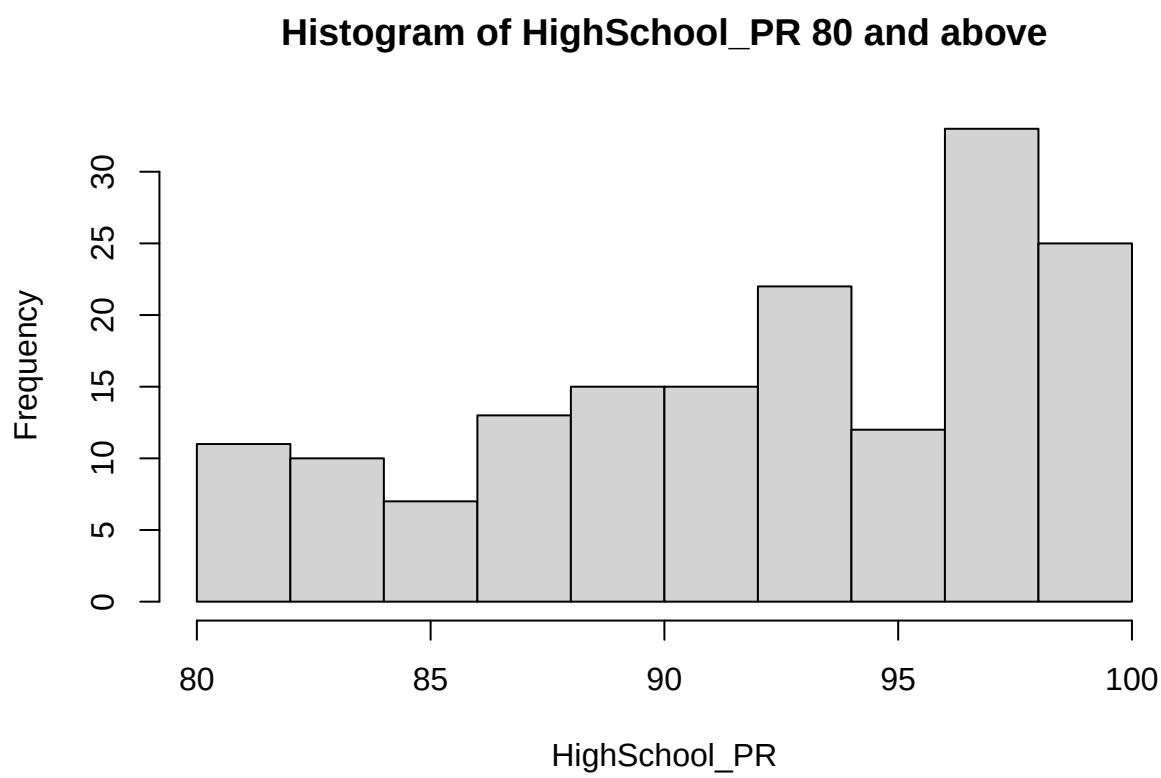
```
## [1] 163
```

```
table(HS_PR_seg)
```

```
## HS_PR_seg
## 80 81 82 83 84 85 86 87 88 89 90 91 92 93 94 95 96 97 98 99
##  3  4  4  4  6  4  3  7  6  8  7  6  9 11 11  7  5 18 15 25
```

Or if you prefer a histogram, we can also create one.

```
hist(HS_PR_seg, xlab="HighSchool_PR",
     main="Histogram of HighSchool_PR 80 and above")
```




Therefore, we create the breakdown of the **HighSchool_PR** ranges: 80-89, 90-94, 95-99. There are 49 records in the 80-89 range, 44 records in 90-94, and 70 records in 95-99. We divided the 90-99 range into 90-94 and 95-99, but the number of **HighSchool_PR** records in the 95-99 range is still higher than any of the other two categories.

However, it is not a good idea to further divide the 95-99 range into 95-97 and 98-99, due to the lack of

geographic information. In Taipei, the high school enrollment is extremely competitive. Students with **HighSchool_PR** 95 and those with **HighSchool_PR** 99 would get admitted to high schools of different rankings.[67] But in other parts of Taiwan, most students with **HighSchool_PR** at least 95 would already qualify for the top local high school, and some rural parts even require a lower **HighSchool_PR** to get into the county's top high school.

```
HS80to89 = length(which(HS_PR_seg >= 80 & HS_PR_seg <= 89))
HS90to94 = length(which(HS_PR_seg >= 90 & HS_PR_seg <= 94))
HS95to99 = length(which(HS_PR_seg >= 95 & HS_PR_seg <= 99))
```

```
print(paste("HighSchool_PR 80-89:",HS80to89))
```

```
## [1] "HighSchool_PR 80-89: 49"
```

```
print(paste("HighSchool_PR 90-94:",HS90to94))
```

```
## [1] "HighSchool_PR 90-94: 44"
```

```
print(paste("HighSchool_PR 95-99:",HS95to99))
```

```
## [1] "HighSchool_PR 95-99: 70"
```

## 6.2 College Entrance Exam Scores at least 60

Similar to the previous section, we also show the frequency of each **College_Score** value that is at least 60. The total counts is 128, fewer than the 163 counts with **HighSchool_PR** 80 or above. The distribution is relatively uniform for **College_Score** values between 60 and 73, with a steep decline in the counts of **College_Score** 74 and 75 (max possible score). On the college entrance exam, only four respondents scored 74 and two scored 75. According to the historical statistics of the college entrance exam in Taiwan, **College_Score** 74 and 75 account for approximately 0.2% of all test takers each year, which is quite a small percentage.

```
CS_Score_seg = data_corr$College_Score[which(data_corr$CS_60up == TRUE)]
length(CS_Score_seg)
```

```
## [1] 128
```

```
table(CS_Score_seg)
```

```
## CS_Score_seg
## 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75
## 10  7  4 10  5 11 12  9  9  6 10  9 12  8  4  2
```

Before we display the histogram, let's create a table to (approximately) convert **College_Score** into PR (Percentile Rank) using 2001-2014 data.[68] This gives readers an understanding of what percentage of test takers (high school students in grade 12) get what scores. For example, a **College_Score** of 70 would be at the 98.5th percentile, i.e., PR 98.5.

```
college_score = c(60,65,70,74,75)
college_pr = c(88, 95, 98.5, 99.9, 100)

data.frame(college_score, college_pr)
```

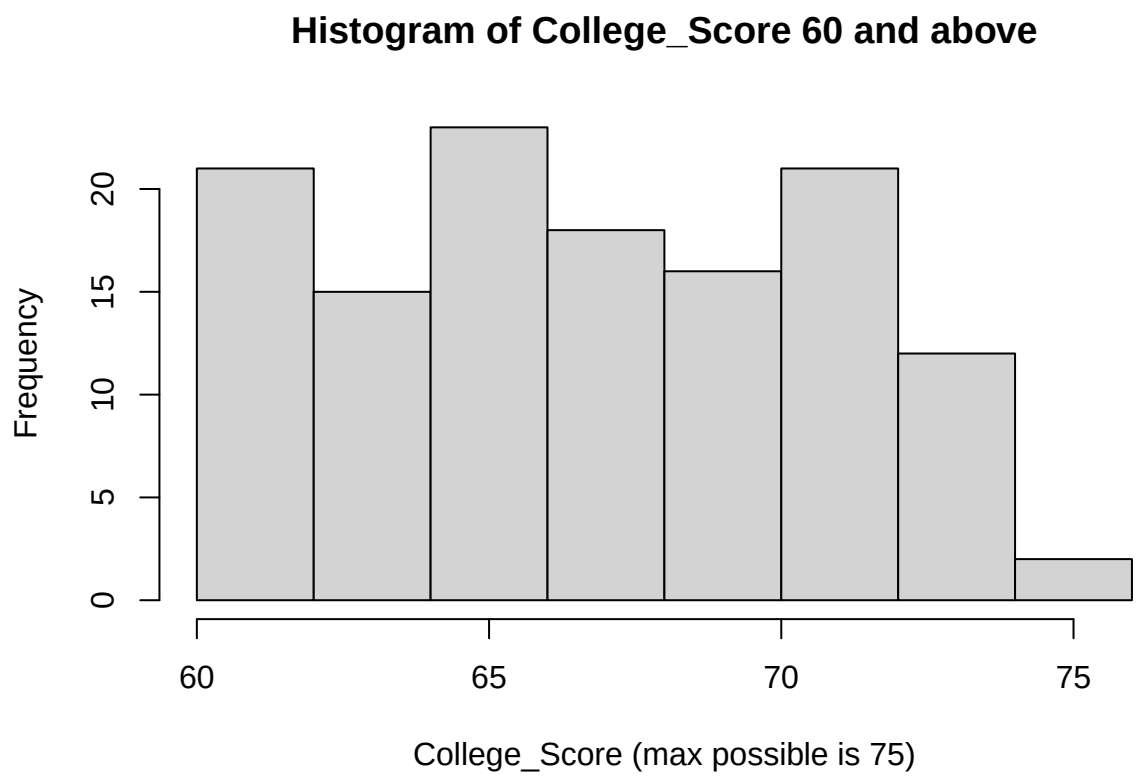
```
##   college_score college_pr
## 1            60       88.0
## 2            65       95.0
## 3            70       98.5
```

[67] https://w199584.pixnet.net/blog/post/28321887
[68] https://web.archive.org/web/20150207042900/http://www.ceec.edu.tw/AbilityExam/AbilityExamStat.htm

```
## 4            74       99.9
## 5            75      100.0
```

Here is the histogram of the **College_Score** values 60 and above.

```
hist(CS_Score_seg, xlab="College_Score (max possible is 75)",
     main="Histogram of College_Score 60 and above")
```




We also create the breakdown of the **College_Score** ranges: 60-64, 65-69, 70-75. There are 36 records in the 60-64 range, 47 records in 65-69, and 45 records in 70-75. This is also relatively more uniform than the **HighSchool_PR** breakdown (49, 44, and 70 records each).

```
CS60to64 = length(which(CS_Score_seg >= 60 & CS_Score_seg <= 64))
CS65to69 = length(which(CS_Score_seg >= 65 & CS_Score_seg <= 69))
CS70to75 = length(which(CS_Score_seg >= 70 & CS_Score_seg <= 75))
```

```
print(paste("College_Score 60-64:",CS60to64))
```

```
## [1] "College_Score 60-64: 36"
```

```
print(paste("College_Score 65-69:",CS65to69))
```

```
## [1] "College_Score 65-69: 47"
```

```
print(paste("College_Score 70-75:",CS70to75))
```

```
## [1] "College_Score 70-75: 45"
```

## 6.3 Bivariate Exploration of Top Scorers

Section 4.3 explored the bivariate relationship between **HighSchool_PR** and **College_Score**, and this time we would like to focus on the high scorers: respondents with **HighSchool_PR** at least 80 and/or **College_Score** at least 60. There are 163 respondents with **HighSchool_PR** 80 or higher, 128 respondents with **College_Score** 60 or higher, and 120 respondents with both. Since the number of respondents with **HighSchool_PR** at least 80 does not equal to the number of respondents with **College_Score** at least 60, we should consider the full 188 records in the data. Hence we add the range 0-79 to **HighSchool_PR**, and the range 0-59 for **College_Score**. We would like to create a 4x4 table for the following ranges:

- **HighSchool_PR** ranges: 0-79, 80-89, 90-94, 95-99
- **College_Score** ranges: 0-59, 60-64, 65-69, 70-75

Here, we use the `for` loop and `if-else` logic to map **HighSchool_PR** and **College_Score** into their corresponding ranges. The `else if` statement is executed when and only when the `if` statement is not true, so we can assign the score to the appropriate category using sequential `if-else` statements for range boundaries.

```
data_corr$HS_range = "set"
data_corr$CS_range = "set"

for(ii in 1:nrow(data_corr)) {
        # High School PR categories
        if (data_corr$HighSchool_PR[ii] <= 79) {
                data_corr$HS_range[ii] = "0-79"
        } else if (data_corr$HighSchool_PR[ii] <= 89) {
                data_corr$HS_range[ii] = "80-89"
        } else if (data_corr$HighSchool_PR[ii] <= 94) {
                data_corr$HS_range[ii] = "90-94"
        } else {
                data_corr$HS_range[ii] = "95-99"
        }

        # College Score Categories
        if (data_corr$College_Score[ii] <= 59) {
                data_corr$CS_range[ii] = "0-59"
        } else if (data_corr$College_Score[ii] <= 64) {
                data_corr$CS_range[ii] = "60-64"
        } else if (data_corr$College_Score[ii] <= 69) {
                data_corr$CS_range[ii] = "65-69"
        } else {
                data_corr$CS_range[ii] = "70-75"
        }
}
```

We continue to use the `R` function `table` to create the 4x4 contingency table between the ranges of **HighSchool_PR** and **College_Score**. As we can see in the horizontal rows, the majority of respondents with **HighSchool_PR** less than 80 have a **College_Score** less than 60. For the respondents with **HighSchool_PR** between 80 and 94, the **College_Score** varies widely. The respondents with **HighSchool_PR** 95 or above performed the best in terms of **College_Score** – most of them scored 65 or higher.

In the vertical columns, the respondents with **College_Score** less than 60 mostly had a **HighSchool_PR** 94 or below; few came from the group of **HighSchool_PR** 95-99. For the respondents with **College_Score** between 60 and 64, their **HighSchool_PR** varied widely. Approximately half of the respondents with **College_Score** had **HighSchool_PR** 95 or above. For the top group of **College_Score** 70 or above, more than three quarters (34 out of 45) came from the respondents with **HighSchool_PR** 95 or higher.

```r
table_4x4 = table(data_corr$HS_range, data_corr$CS_range,
                  dnn=c("HighSchool_PR","College_Score"))
table_4x4
```

```
##              College_Score
## HighSchool_PR 0-59 60-64 65-69 70-75
##         0-79    17     4     0     4
##         80-89   19    16    10     4
##         90-94   18     9    14     3
##         95-99    6     7    23    34
```

The contingency table seems to show a positive association between **HighSchool_PR** and **College_Score**, because most counts are in the top-left and bottom-right. Respondents with a good **HighSchool_PR** score are more likely to achieve a good **College_Score**, but this is not guaranteed. Respondents who scored poorly in **HighSchool_PR** still had a small chance to do exceptionally well in **College_Score** later. Our observations align with the conventional wisdom that **HighSchool_PR** and **College_Score** are somewhat related, but a high score on **HighSchool_PR** does not guarantee a high score on **College_Score**.

### 6.3.1 Hypothesis Testing Framework

We perform hypothesis testing to statistically validate the positive association, and the hypotheses are:

- $H_0$ (null hypothesis): No association exists between the categorical counts of **HighSchool_PR** and **College_Score**.
- $H_1$ (alternative hypothesis): An association exists between the categorical counts of **HighSchool_PR** and **College_Score**.

Note that we set up a two-sided hypothesis test because we are interested in the association between the two variables, no matter it is positive or negative.[69]

The **p-value** measures statistical significance, and it is **the probability of observing something at least as extreme as the data, given the assumption that $\mathbf{H_0}$ is true** (Diez et al., 2019).

- When p-value $> 0.05$, we fail to reject $H_0$ and conclude that $H_0$ is true.
- When p-value $< 0.05$, we reject $H_0$ and conclude that $H_1$ is true.

In the case that p-value $< 0.05$, we say that the observed difference is statistically significant. But note that we do not know the underlying truth and we can make mistakes. We may make one of the two mistakes in hypothesis testing:

- Type 1 error: $H_0$ is true but we reject $H_0$.
- Type 2 error: $H_0$ is false but we fail to reject $H_0$.

The good thing is that with p-value $< 0.05$, we limit the chances of making a Type 1 error to be less than 5%. This threshold is also called the significance level.[70] When the cost of making a Type 1 error is higher, we can require p-value $< 0.01$ or even 0.001 to reject $H_0$ instead.

The definition of p-value is often confused with the probability of the null hypothesis being true, which is incorrect (Goodman, 2008). P-values have been widely used and misused in scientific research, up to the extent that the American Statistical Association (2016) released a statement on the proper use and interpretation of the p-values. Ideally we should consider the whole research design and the results' practical importance, rather than make conclusions solely based on statistical significance (Wasserstein et al., 2019).

**Remark**: Bayesian hypothesis testing (Kruschke and Liddell, 2018) outputs real probability values because the whole framework is generated using probability distributions. The Bayesian 95% credible intervals are 95% posterior probabilities, as opposed to the 95% confidence intervals in the frequentist approach, which do not have 95% probability.

[69] https://statisticsbyjim.com/hypothesis-testing/one-tailed-two-tailed-hypothesis-tests/
[70] https://www.statsdirect.com/help/basics/p_values.htm

### 6.3.2 Pearson's Chi-Squared Test

The most common form of hypothesis testing is to check whether a coefficient is zero or not in linear regression, but this is by no means the only form. For categorical data, the Pearson's chi-squared test[71] can be used to evaluate whether the categorical counts of **HighSchool_PR** and **College_Score** are due to random chance. The test statistic $\chi^2$ (chi-squared) is defined as below, and it asymptotically approaches a $\chi^2$ distribution.

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \tag{1}$$

There are $n$ cells in the table. For each cell $i$, $O_i$ is the number of observations in that cell and $E_i$ is the number of expected counts under the null hypothesis. Each $E_i$ is conditioned on row and column totals (i.e., marginal totals),[72] so $E_i$ differs for each cell $i$. For example, we have an unfair coin which lands heads 80% of the time, and we toss the coin 100 times. Then $E_{\text{heads}}$ is 80 and and $E_{\text{tails}}$ is 20.

Then the test statistic $\chi^2$ is compared with a $\chi^2$ distribution to calculate the p-value, given the degrees of freedom. When we generate $m$ independent scores from a random sample, we have $m - 1$ degrees of freedom because the $m$ scores are restricted by their sample mean. When we have $a$ rows and $b$ columns in the table, we have $(a - 1)(b - 1)$ degrees of freedom because each row and each column are restricted to the subtotal, on behalf of the grand total (total number of observations in the data).

It is tedious to calculate the Pearson's chi-squared test by hand, so we use the function `chisq.test` in the `R` package `stats`. But we should still try to understand the rationale behind the `chisq.test` function to verify that we are using it correctly. For example, the test statistic $\chi^2$ has `df = 9` degrees of freedom, because `df` is computed as $(4 - 1) \times (4 - 1)$.

The results show p-value $< 0.05$, so we reject $H_0$ and conclude that a positive association exists between the categorical counts of **HighSchool_PR** and **College_Score**. This is statistical evidence to support our observation in the contingency table.

```
results = chisq.test(table_4x4)
```

```
## Warning in chisq.test(table_4x4): Chi-squared approximation may be incorrect
```

```
results
```

```
##
##  Pearson's Chi-squared test
##
## data:  table_4x4
## X-squared = 69.98, df = 9, p-value = 1.536e-11
```

**Remark**: We can add `simulate.p.value = TRUE` to suppress the warning `Chi-squared approximation may be incorrect`, so the p-values would be simulated. But given the small cell sizes, the estimates are inaccurate anyway. In fact, `simulate.p.value = TRUE` uses simulation conditional on the marginal totals, so `chisq.test` performs the Fisher's exact test[73] with the multivariate version of hypergeometric distributions.

Let's take a look at what `chisq.test` provides us, and verify that we are using the function correctly. We should always leverage a package to do the computation when there is one that can do the job, but we also need to know exactly what we are doing to reap the results.

The `chisq.test` reads in the observed counts, i.e., the data in `table_4x4`, as shown earlier in Section 6.3.

```
observed = results$observed
observed
```

[71] https://data-flair.training/blogs/chi-square-test-in-r/
[72] https://www.stats4stem.org/chi-square-test-for-homogeneity
[73] https://mathworld.wolfram.com/FishersExactTest.html

```
##              College_Score
## HighSchool_PR 0-59 60-64 65-69 70-75
##         0-79    17     4     0     4
##         80-89   19    16    10     4
##         90-94   18     9    14     3
##         95-99    6     7    23    34
```

The expected counts are simulated and conditioned on row and column totals, so each cell has a different number of expected counts. For example, the first row for **HighSchool_PR** 0-79 adds up to 25, which is the same in both `observed` and `expected` tables. The first column for **College_Score** 0-59 adds up to 60, and the sum is also the same in both tables.

```
expected = results$expected
expected
```

```
##              College_Score
## HighSchool_PR      0-59     60-64 65-69     70-75
##         0-79   7.978723  4.787234  6.25  5.984043
##         80-89 15.638298  9.382979 12.25 11.728723
##         90-94 14.042553  8.425532 11.00 10.531915
##         95-99 22.340426 13.404255 17.50 16.755319
```

We cannot just assume $E_i = 188/16 = 11.75$ given 188 total records and 16 cells, because this does not adhere to the original row and column totals. Setting all cells to have equal expected counts means they are conditioned on only the grand total (188 records), so the degrees of freedom is $16 - 1 = 15$. With a 4x4 table, the degrees of freedom should be $(4 - 1) \times (4 - 1) = 9$.

Let's use Equation (1) to compute and verify the test statistic $\chi^2$, and we get the same result as the `chisq.test` function. The manual calculation is for educational purposes only, and readers do not have to do this in a real-time project.

```
test_stat = 0

for (ii in 1:nrow(observed)) {
  for (jj in 1:nrow(observed)) {
    test_stat = test_stat + ((observed[ii,jj] - expected[ii,jj])^2)/expected[ii,jj]
  }
}

test_stat
```

```
## [1] 69.98023
```

# 7 Logistic Regression

We decided to perform a different model to quantify the relationship between **HighSchool_PR** and **College_Score**, because we concluded in Chapter 5 that it is inappropriate to perform an ordinary linear regression. (We also tried the square root transformation, but it did not work out, either.) Let's try another statistical model to evaluate the relationship between the two variables. Typically when the ordinary linear regression model is ruled out, the next candidate is a generalized linear model, such as logistic regression and Poisson regression. Choosing a different model may involve modifying the details of the problem statement. The original problem statement is to investigate the relationship between **College_Score** and **HighSchool_PR**, but it is flexible and does not require the relationship to be linear.

We would like to redefine the problem statement to "estimate the probability of **College_Score** at least 65, given the student's **HighSchool_PR**." Since the variance of **College_Score** depends on **HighSchool_PR**, the assumptions of linear regression are violated, making linear regression an inappropriate model. We

decided to focus on whether **College_Score** is at least a particular value instead, so the response variable is binary. We selected 65 as the cutoff point for **College_Score** because this is close to the median, making the number of values about the same in the two categories. We would like to balance the number of datapoints in each category of the response variable.

Logistic regression is a generalized linear model that uses a binary response variable, and the equation models the probability of an event occurring or not. That's why we set up the new problem statement this way, and Section 7.1 gives a brief introduction to the logistic regression model. Although logistic regression is not typically covered at the Statistics 101 level, we would like to give the readers a head start of generalized linear models. We explained the requirements of linear regression in Chapter 5, and now we would like to expand the regression model to additional forms. We decided to introduce generalized linear models early on, because we would like to show that there exist methods to analyze the data `ptt_SENIORHIGH_data.csv` other than linear regression.

We try to explain the logistic regression in basic terminology. But if the readers feel like they cannot understand the mathematics, it is totally fine to skip this chapter and move on to Chapter 8 for model validation. The model evaluation concepts are not related to the logistic regression itself, so understanding the model details is not always a prerequisite. We can regard the model as a blackbox, which simply produces the binary classification output.

## 7.1 Brief Introduction of Logistic Regression

Logistic regression is used to model categorical outcomes, especially a binary response variable. For example, if the response variable is whether an event $Y$ occurs or not, then it can only have two values – not occurred (0) and occurred (1). We model the probability that the event $Y$ occurred, denoted as $p = P(Y = 1)$, and it is between 0 and 1. But in a linear regression $y = \alpha + \beta x$, the response variable $y$ can be any real number. Therefore, we transform the probability $p$ to the log odds $\log(\frac{p}{1-p})$, so that its range spans the whole real line like $y$. Note that the odds $\frac{p}{1-p}$ is always positive, as long as $p$ is not exactly 0 or 1.

The equation for logistic regression is written as below:

$$\text{logit}(p) = \log(\frac{p}{1-p}) = \alpha + \beta X$$

The notation is similar to a linear regression, but the interpretation is slightly different. $\alpha$ is the intercept, and $\beta$ is the coefficient. When $\beta$ increases by one unit, the log odds $\log(\frac{p}{1-p})$ increases by $\beta$ units. In other words, when $\beta$ increases by one unit, the odds $\frac{p}{1-p}$ are multiplied by $e = \exp(1) \approx 2.71828$.

The probability $p$ can be estimated from the logistic regression model with the equation below.

$$p = \frac{\exp(\alpha + \beta X)}{1 + \exp(\alpha + \beta X)}$$

The intercept $\alpha$ serves as a baseline when $X = 0$, and the probability $p$ can be estimated by $p = \frac{\exp(\alpha)}{1 + \exp(\alpha)}$. Just like in an ordinary linear regression, the intercept may or may not have practical meaning. For example, if we examine the relationship between people's height and weight, nobody is going to have height of 0 inches.

For the advanced readers, we recommend reading the textbook *Categorical Data Analysis* (Agresti, 2003) to learn more about logistic regression and other generalized linear models for categorical data.

## 7.2 Estimate the Probability of Scoring Well on the College Entrance Exam

We would like to redefine the problem statement to "estimate the probability of **College_Score** at least 65, given the student's **HighSchool_PR**." Since the variance of **College_Score** depends on **HighSchool_PR**, the assumptions of linear regression are violated, making linear regression an inappropriate model. That's why we decided to focus on whether **College_Score** is at least a particular value instead, so the response variable is binary. We selected 65 as the cutoff point for **College_Score** because this is close to the median, making the number of values about the same in the two categories. We would like to balance the number of datapoints in each category of the response variable.

```r
print(paste("College_Score at least 65:", sum(data_corr$College_Score >= 65)))
```

```
## [1] "College_Score at least 65: 92"
```

```r
print(paste("College_Score below 65:", sum(data_corr$College_Score < 65)))
```

```
## [1] "College_Score below 65: 96"
```

The event $Y$ we would like to observe is getting a **College_Score** at least 65. We define

$$p = P(Y = 1) = P(\textbf{College_Score} \geq 65),$$

i.e., the probability of getting a **College_Score** at least 65. And the independent variable $X$ is the **HighSchool_PR**, whose values are between 0 and 99.

Then we implement the logistic regression with the equation

$$\text{logit}(p) = \log(\frac{p}{1-p}) = \alpha + \beta X.$$

The model is written as `glm(y ~ x, data=data, family="binomial")` in `R` code, where `glm` stands for generalized linear regression. The `family` option is set to binomial, because the response variable is binary and can only have values 0 or 1. Hence our logistic model can be written as

$$\text{logit}(P(\textbf{College_Score} \geq 65)) \sim \alpha + \beta * \textbf{HighSchool_PR}.$$

Let's create the logistic regression model in `R` as below.

```r
# 1. Create the binary variable first.
# 2. model = glm( y ~ x, data=data, family="binomial")
# 3. summary(model)
# https://stats.idre.ucla.edu/r/dae/logit-regression/

data_corr$CS_65up = data_corr$College_Score >= 65
model = glm(CS_65up ~ HighSchool_PR, data=data_corr, family="binomial")
summary(model)
```

```
##
## Call:
## glm(formula = CS_65up ~ HighSchool_PR, family = "binomial", data = data_corr)
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -13.63939    2.45755  -5.550 2.86e-08 ***
## HighSchool_PR   0.14993    0.02674   5.607 2.06e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 260.54  on 187  degrees of freedom
## Residual deviance: 211.98  on 186  degrees of freedom
## AIC: 215.98
##
## Number of Fisher Scoring iterations: 5
```

## 7.3 Model Interpretation: Point Estimates

The `Pr(>|z|)` is the p-value of each independent variable, and we can see that **HighSchool_PR** is statistically significant because p-value $< 0.05$. When **HighSchool_PR** increases by one, the log odds $\log(\frac{p}{1-p})$ of getting **College_Score** at least 65 increases by approximately 0.15. After transforming log odds $\log(\frac{p}{1-p})$ back to odds $\frac{p}{1-p}$, we get that the odds are multiplied by $e = \exp(0.15) \approx 1.161$. In other words, when **HighSchool_PR** increases by one, the odds of getting **College_Score** at least 65 increases by approximately 16.1%.

For better reproducibility of coefficients, these can be retrieved using the code below.

```r
model$coefficients
```

```
##   (Intercept) HighSchool_PR
##   -13.6393909     0.1499257
```

We can also round the coefficients to the third digit after the decimal point. But for better precision, we do not recommend rounding numbers until we reach the final calculation results.

```r
round(model$coefficients, digits=3)
```

```
##   (Intercept) HighSchool_PR
##       -13.639         0.150
```

Here is the exponential of the coefficients, because we need to transform log odds into odds.

```r
exp(model$coefficients)
```

```
##   (Intercept) HighSchool_PR
##  1.192581e-06  1.161748e+00
```

The intercept serves as a baseline for the logistic regression model when **HighSchool_PR** is 0. We can predict $p$ under this condition, and find out how likely this (fictitious) person is going to get **College_Score** at least 65. The estimated probability is extremely low, less than 0.01%. Nevertheless, the value of **HighSchool_PR** starts at 1, so the intercept does not have an intrinsic meaning. (And typically most students who score less than 10 in **HighSchool_PR** would not be interested in attending college, either.)

```r
alpha = as.numeric(model$coefficients[1])
beta = as.numeric(model$coefficients[2])

p_intercept = exp(alpha)/(1+exp(alpha))
p_intercept
```

```
## [1] 1.192579e-06
```

In comparison, when **HighSchool_PR** is 99, the model estimates that the student has a 76.9% chance to achieve a **College_Score** of 65 or higher.

```r
p_pr99 = exp(alpha + beta*99)/(1+exp(alpha + beta*99))
p_pr99
```

```
## [1] 0.7691025
```

If we look at the data, there are 25 respondents with **HighSchool_PR** 99, and only one of them scored below 65 in the **College_Score**. The data show the conditional probability to be 96%.

$$P(\textbf{College_Score} \geq 65 | \textbf{HighSchool_PR} = 99) = \frac{24}{25} = 96\%.$$

In this **HighSchool_PR** 99 group, more than half of the respondents (14, to be exact) had a **College_Score** between 71 and 73. Nevertheless, **HighSchool_PR** 99 is not an (almost) necessary condition to achieve **College_Score** 65 or higher. In the group of **College_Score** 65 or higher, only 24 out of the 92 respondents had **HighSchool_PR** 99, which is less than a quarter.

$$P(\textbf{HighSchool_PR} = 99 | \textbf{College_Score} \geq 65) = \frac{24}{92} \approx 26\%.$$

```r
num_pr99 = sum(data_corr$HighSchool_PR == 99)
num_cs65 = sum(data_corr$College_Score >= 65)
num_pr99_and_cs65 = sum(data_corr$HighSchool_PR == 99 & data_corr$College_Score >= 65)
```

```r
print(paste("Number of respondents with HighSchool_PR 99:",num_pr99))
```

```
## [1] "Number of respondents with HighSchool_PR 99: 25"
```

```r
print(paste("Number of respondents with College_Score 65 or better:",num_cs65))
```

```
## [1] "Number of respondents with College_Score 65 or better: 92"
```

```r
print(paste("Number of respondents with HighSchool_PR 99 and College_Score 65 or better:",
            num_pr99_and_cs65))
```

```
## [1] "Number of respondents with HighSchool_PR 99 and College_Score 65 or better: 24"
```

```r
sort(data_corr$College_Score[which(data_corr$HighSchool_PR==99)])
```

```
##  [1] 57 65 66 67 67 68 68 70 70 71 71 71 71 71 72 72 72 72 73 73 73 73 73 74 75
```

## 7.4 Model Interpretation: 95% Confidence Intervals

In addition to the point estimates, we also need to provide the corresponding 95% confidence intervals to account for uncertainty. The lower bound is the 2.5th percentile, and the upper bound is the 97.5th percentile. Statistical significance at 5% means that the 95% confidence interval does not include 0. Let's start with the intercept and the coefficient for **HighSchool_PR** using the `R` function `confint`. Neither the intercept nor the coeffient's confidence interval includes 0, so both are statistically significant.

Although the intercept $\alpha$'s confidence interval seems wide, the exponential version $\exp(\alpha)$ is extremely small for both ends. Especially when the intercept (baseline for **HighSchool_PR** $= 0$) does not have practical meaning in this data, we do not need to be overly concerned about the intercept.

On the other hand, the coefficient $\beta$ has a point estimate of approximately 0.15, with a 95% confidence interval [0.101, 0.206]. When **HighSchool_PR** increases by one, we can expect an increase between 0.101 and 0.206 in the log odds $\log(\frac{p}{1-p})$ of getting **College_Score** at least 65. We can also transform log odds $\log(\frac{p}{1-p})$ back to odds $\frac{p}{1-p}$, and get an expected increase factor between 1.106 and 1.229. We are 95% confident that the odds of getting **College_Score** at least 65 would increase by between 10.6% and 22.9%, given that **HighSchool_PR** increases by one.

```r
confint(model)
```

```
## Waiting for profiling to be done...
```

```
##                     2.5 %     97.5 %
## (Intercept)   -18.7973934 -9.1378744
## HighSchool_PR   0.1008108  0.2059154
```

```
exp(confint(model))
```

```
## Waiting for profiling to be done...
```

```
##                      2.5 %       97.5 %
## (Intercept)   6.861132e-09 0.0001075156
## HighSchool_PR 1.106067e+00 1.2286493088
```

We can also calculate the 95% confidence interval for the predicted probability of getting **College_Score** at least 65, given that the respondent scored a 99 on **HighSchool_PR**. However, the confidence interval is [0.00012, 0.99999], which does not make practical sense because a probability has natural boundaries of [0,1].

Why is this interval so wide? The linear component is $\alpha + \beta X$, so the range depends on the value of **HighSchool_PR**. When **HighSchool_PR** is large (say, 99), the 95% confidence interval of $\alpha + \beta X$ becomes extremely wide.

```
ci_matrix = confint(model)
alpha_ci = ci_matrix[1,]
beta_ci = ci_matrix[2,]
p_pr99_ci = exp(alpha_ci + beta_ci*99)/(1+exp(alpha_ci + beta_ci*99))
p_pr99_ci
```

```
##        2.5 %       97.5 %
## 0.0001481513 0.9999869636
```

The interval would be narrow for a small value of **HighSchool_PR**, such as less than 10. But people with extremely low **HighSchool_PR** are unlikely to be interested in taking the college entrance exam at all. Hence we are not going to calculate the 95% confidence interval for small **HighSchool_PR** values.

**Remark**: The readers may wonder how the author(s) remember all of the `R` commands for the model. In fact, we don't need to! We can use the function `objects` to find all available outputs from the model, and the object names are descriptive.

```
objects(model)
```

```
##  [1] "aic"               "boundary"          "call"
##  [4] "coefficients"      "contrasts"         "control"
##  [7] "converged"         "data"              "deviance"
## [10] "df.null"           "df.residual"       "effects"
## [13] "family"            "fitted.values"     "formula"
## [16] "iter"              "linear.predictors" "method"
## [19] "model"             "null.deviance"     "offset"
## [22] "prior.weights"     "qr"                "R"
## [25] "rank"              "residuals"         "terms"
## [28] "weights"           "xlevels"           "y"
```

## 7.5 Overall Model Results

In addition to the coefficient estimates, we also need to perform model validation to examine how well the model fits the data. Let's start with visualizing the predicted probability of getting **College_Score** at least 65 and the **HighSchool_PR** values in the data. The former can be obtained from the `fitted.values` object of the model. The highest predicted probability occurs at **HighSchool_PR** 99, but the predicted

probability of getting **College_Score** at least 65 is still less than 80%. (So high school students should still study hard for the college entrance exam, despite getting an excellent **HighSchool_PR** score.)

```
print(paste("The highest predicted probability is",max(model$fitted.values)))
```

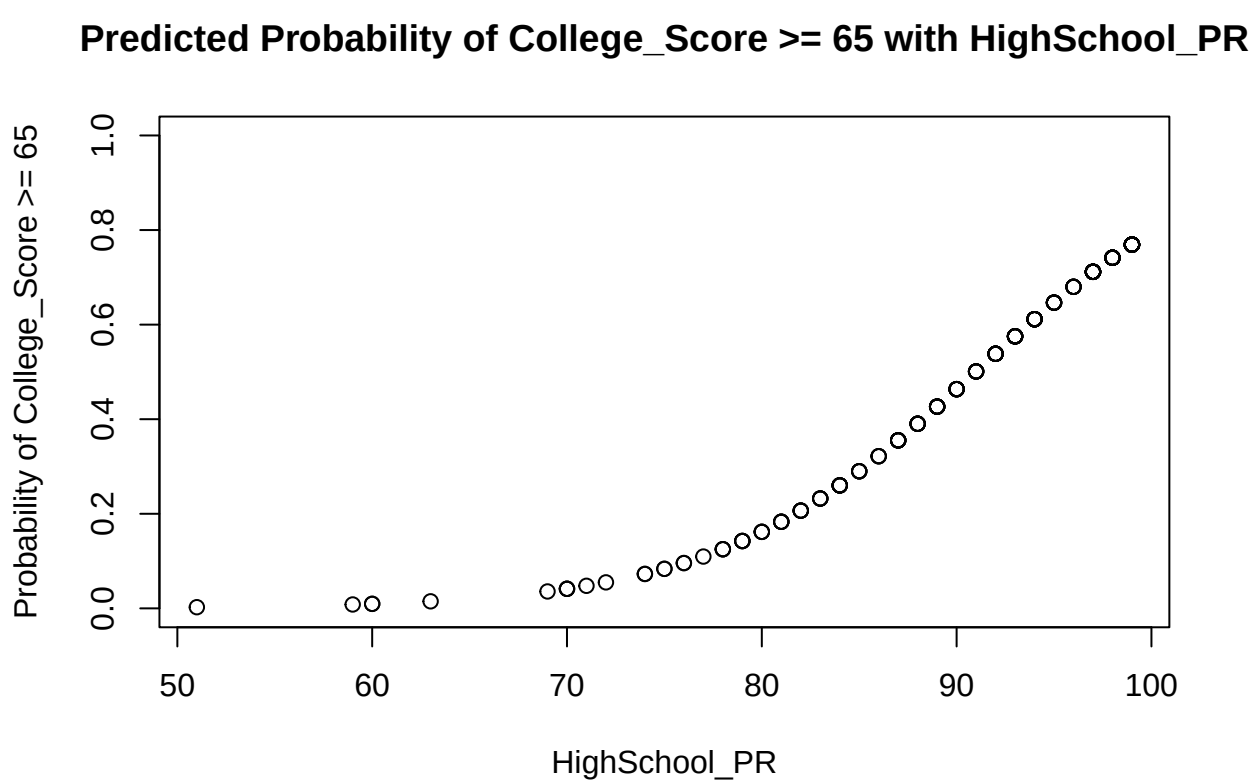
```
## [1] "The highest predicted probability is 0.769102549200382"
```

```
print(paste("This occurs at HighSchool_PR",
      data_corr$HighSchool_PR[which.max(model$fitted.values)]))
```

```
## [1] "This occurs at HighSchool_PR 99"
```

The graph shows different trends for different segments of the **HighSchool_PR**. When **HighSchool_PR** is less than 70, the predicted probability of getting **College_Score** at least 65 is close to zero. But we should take this observation with caution, because we have few data points with **HighSchool_PR** less than 70. When **HighSchool_PR** is between 70 and 79, the predicted probability increases with an almost linear trend. Starting at **HighSchool_PR** 80, the predicted probability increases with a steeper slope. Finally, the growth of the predicted probability slows down when **HighSchool_PR** reaches 98 (the maximum possible **HighSchool_PR** is 99).

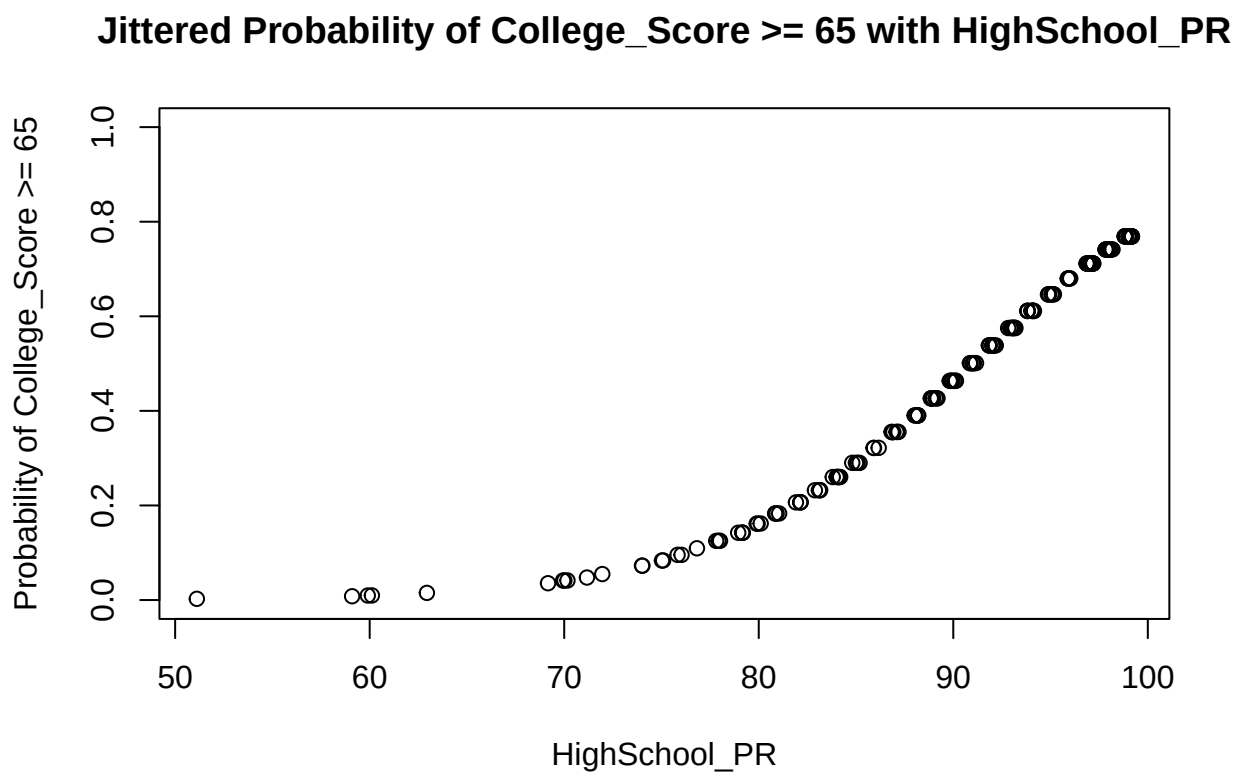
```
# First version of graph
yy = model$fitted.values
xx = data_corr$HighSchool_PR
plot(xx, yy, ylim=c(0,1),
     main="Predicted Probability of College_Score >= 65 with HighSchool_PR",
     xlab="HighSchool_PR",
     ylab="Probability of College_Score >= 65")
```




Since there are many repetitive values in **HighSchool_PR**, let's apply the `jitter` function to add random

noise to the data for display. (Remember to set a random seed for reproducibility.) We can see that the deeper black circles indicate more records in the data, which are concentrated in the higher **HighSchool_PR** values.

```
# Add jittering because of many repetitive values in HighSchool_PR
set.seed(21)
new_xx = jitter(xx)
new_yy = jitter(yy)
plot(new_xx, new_yy, ylim=c(0,1),
     main="Jittered Probability of College_Score >= 65 with HighSchool_PR",
     xlab="HighSchool_PR",
     ylab="Probability of College_Score >= 65")
```




# 8 Model Validation: In-Sample Prediction

We built the logistic regression model in Chapter 7, and now it is time for model validation, i.e., evaluate the model performance. We will focus on machine learning concepts in this chapter, and the readers simply need to know that **the model does binary classification**. For more about machine learning, we recommend the book *Introduction to Machine Learning* (Alpaydin, 2020). It explains a wide range of machine learning algorithms and applications, including the recent advances in deep learning and neural networks.

We start with in-sample prediction to obtain the binary classification results, then we explain how to interpret the 2x2 confusion matrix output. There are two actual outcomes for **College_Score** – at least 65 or not. There are also two predicted outcomes for **College_Score**, and we compare the predicted outcomes with the actual ones. Next, we would like to examine the model performance for different **HighSchool_PR** scores, in order to identify whether the model does better in higher or lower **HighSchool_PR**, or performs about

the same. In Chapter 9, we will perform out-of-sample prediction on the model; that is, train the model on some data and test it on different data.

## 8.1 Implementation of In-Sample Prediction

Let's start the model validation with in-sample prediction; that is, using the data to predict the outcome of whether **College_Score** is at least 65 for each value of **HighSchool_PR** for values already in the data. If predicted probability of **College_Score** at least 65 (i.e., `fitted.values`) is greater than or equal to 0.5, we assign the predicted value as `TRUE`. We can safely use 0.5 as the probability threshold because the data are quite balanced. In other words, the proportion of **College_Score** at least 65 is close to 0.5 in the data (0.489, to be exact). If the data are imbalanced, say 80% of the records belong to one category, we should adjust the probability threshold in our classification prediction model.

```
# nrow(data_corr) # 188
# sum(data_corr$CS_65up) # 92
# sum(data_corr$CS_65up)/nrow(data_corr) # 0.489

print(paste("There are",nrow(data_corr),"records in the data, with",
            sum(data_corr$CS_65up),
            "of them have College_Score at least 65."))
```

```
## [1] "There are 188 records in the data, with 92 of them have College_Score at least 65."
```

```
print(paste("This is a proportion of",
            round(sum(data_corr$CS_65up)/nrow(data_corr),digits=3),
            ", which is close to 0.5."))
```

```
## [1] "This is a proportion of 0.489 , which is close to 0.5."
```

Let's create a confusion matrix to show the comparison between the actual outcomes and predicted outcomes, i.e., whether each respondent obtained **College_Score** at least 65 or not. Note that a confusion matrix is slightly different than the contingency table in Section 5.2, although both record counts in a matrix. A confusion matrix involves the predicted results, while a contingency table simply observes the categories in the data.

```
# Data
actual_65up = data_corr$CS_65up

# Predicted results
predicted_65up = model$fitted.values >= 0.5

# Confusion matrix
confusion = table(actual_65up, predicted_65up)
# revert the order of FALSE and TRUE
confusion = confusion[2:1, 2:1]
confusion
```

```
##            predicted_65up
## actual_65up TRUE FALSE
##       TRUE    72    20
##       FALSE   35    61
```

Below is the percentage version of the confusion matrix.

```
prop.table(confusion)
```

```
##            predicted_65up
## actual_65up      TRUE     FALSE
```

```
##       TRUE  0.3829787 0.1063830
##       FALSE 0.1861702 0.3244681
```

The table below shows the meaning of each cell of the confusion matrix. Assume that "positive" means getting **College_Score** at least 65, which is equivalent to the `TRUE` label in `actual_65up` and `predicted_65up`. The term "negative" means not getting **College_Score** at least 65, so it is equivalent to the `FALSE` label. For each cell, we have:[74]

- **True Positive (TP)**: The datapoint is actually positive and is predicted as positive, so the model correctly predicts the positive outcome.
- **True Negative (TN)**: The datapoint is actually negative and is predicted as negative, so the model correctly predicts the negative outcome.
- **False Positive (FP)**: The datapoint is actually negative but is predicted as positive, so the model **in**correctly predicts the positive outcome, i.e., false alarm.
- **False Negative (FN)**: The datapoint is actually positive but is predicted as negative, so the model **in**correctly predicts the negative outcome, i.e., error.

```
##                 | Predicted Positive | Predicted Negative
## --------------  | ------------------ | ------------------
## Actual Positive |      True Positive |     False Negative
## --------------- | ------------------ | ------------------
## Actual Negative |     False Positive |      True Negative
```

We can retrieve each element in the confusion matrix by specifying the labels for each row and each column. The row indicates how many respondents actually obtained **College_Score** at least 65, and the column indicates how many respondents were predicted to have the positive outcome. Instead of writing the syntax like `confusion[1,2]`, we write it in a clearer way `confusion["TRUE","FALSE"]`. (Don't make the confusion matrix more confusing!) The readers can easily see that this is the number of respondents who we did not predict to have a **College_Score** at least 65, but they actually did.

```r
# row = actual_65up, column = predicted_65up
tp = confusion["TRUE","TRUE"]
fn = confusion["TRUE","FALSE"]
fp = confusion["FALSE","TRUE"]
tn = confusion["FALSE","FALSE"]

print(paste(tp,fn,fp,tn))
```

```
## [1] "72 20 35 61"
```

We can also verify that the retrieved numbers are exactly the same as in the original confusion matrix.

```r
confusion
```

```
##            predicted_65up
## actual_65up TRUE FALSE
##       TRUE    72    20
##       FALSE   35    61
```

## 8.2 Interpretation of Confusion Matrix

After creating the confusion matrix to record the model performance, we need to interpret the numbers and define metrics to measure the performance. We start with the overall **accuracy** to calculate how many datapoints the model predicted correctly. A datapoint is correctly predicted if one of the two scenarios occurs:

- True Positive: The datapoint is actually positive and it is predicted as positive.
- True Negative: The datapoint is actually negative and it is predicted as negative.

[74] https://developers.google.com/machine-learning/crash-course/classification/true-false-positive-negative

In the context of our dataset, "positive" means getting a **College_Score** 65 or higher, and "negative" means getting a **College_Score** of 64 or lower. We run the model to predict the respondents' college entrance exam outcome given their **HighSchool_PR**. In mathematical terms, accuracy can be calculated as below:

$$\text{Accuracy} = \frac{\text{True Positive} + \text{True Negative}}{\text{Actual Positive} + \text{Actual Negative}} = \frac{\text{True Positive} + \text{True Negative}}{\text{All Datapoints}}$$

Plugging in the numbers from our model, we get

$$\text{Accuracy} = \frac{72 + 61}{72 + 20 + 35 + 61} \approx 70.74\%.$$

Our model correctly predicts whether the **College_Score** would be at least 65 or not around 70% of the time, which is better than a 50-50 coin flip. This means our model is informative, and the results align with the prior knowledge that a higher **HighSchool_PR** is more likely to lead to **College_Score** at least 65.

Note that when the data are imbalanced (say, 98% of the records belong to one category), accuracy is not a good measure of model performance.[75] The model can simply predict all datapoints to be in the larger category, and achieve 98% accuracy. The good news is that we get a relatively balanced dataset by setting the classification threshold of **College_Score** to be 65, as we explained at the beginning of Section 7.2. In fact, 48.9% of the respondents have a **College_Score** of 65 or higher.

```
print(paste("The proportion of respondents with College_Score 65 or higher is",
            round(sum(data_corr$CS_65up)/nrow(data_corr), digits = 3)))
```

```
## [1] "The proportion of respondents with College_Score 65 or higher is 0.489"
```

### 8.2.1 Precision and Recall

**Precision** and **recall** are also two widely-used metrics to measure the performance of the prediction model, and most importantly, they do not depend much on the proportions of data categories. Precision is defined as the percentage of true positives among all datapoints the model predicted to be positive. In our example, precision is the percentage of the respondents who actually got **College_Score** at least 65, among the respondents we predicted to make this achievement given their **HighSchool_PR**.

$$\text{Precision} = \frac{\text{True Positive}}{\text{Predicted Positive}} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}$$

Plugging in the numbers from our model, we get

$$\text{Precision} = \frac{72}{72 + 35} \approx 67.29\%.$$

The precision would be useful if we use the predictive information to decide to invest in which high school students based on their **HighSchool_PR**. If we predict a student to get **College_Score** at least 65, he/she has about a 67.29% chance to make it, which is more than two-thirds. Since there are different tiers of high schools in Taiwan based on **HighSchool_PR**, many resources are given to the top tier high schools, because these students have the highest chance to do well on the college entrance exam. Nevertheless, other social factors also play a role in the overall policy decision-making process. For example, the government may decide to put more resources into remote rural high schools to empower disadvantaged students to succeed, which is beneficial for upward social mobility.

On the other hand, recall is defined as the percentage of model-predicted positives among all datapoints that are actually positive. In our example, recall is the percentage of the respondents we predicted to have **College_Score** at least 65 given their **HighSchool_PR**, among the respondents who actually made this achievement.

[75] https://www.analyticsvidhya.com/blog/2017/03/imbalanced-data-classification/

$$\text{Recall} = \frac{\text{True Positive}}{\text{Actual Positive}} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}$$

Plugging in the numbers from our model, we get

$$\text{Recall} = \frac{72}{72 + 20} \approx 78.26\%.$$

The recall would also be useful if we use the predictive information to decide to invest in which high school students based on their **HighSchool_PR**. We need to remember that among the high school graduates who got a **College_Score** at least 65, only 78.26% of them were predicted to have such potential. In other words, the remaining 21.74% of high school students did better than the predictive model had expected. It is possible that they were smart but accidentally did poorly on the high school entrance exam, and they would have achieved **College_Score** at least 65 regardless. Another possibility is that they did not study much in middle school and got a low **HighSchool_PR**, but they received extra help and/or studied much harder for the college entrance exam to get a **College_Score** at least 65. The lesson is that by pre-selecting students based on **HighSchool_PR**, we would still get some "dark horses", i.e., the students who performed much better on the **College_Score** than we had expected.

### 8.2.2 False Positive Rate and False Negative Rate

**False positive rate** and **false negative rate** are typically used to measure the accuracy of a medical screening test for a disease (NCSS, nd), where "positive" means having the disease and "negative" means not having the disease. In our dataset, "positive" means something much better – getting a **College_Score** 65 or higher. "Negative" means not achieving this.

False positive rate is the probability of an actual negative being classified as a positive, and it is also called a "false alarm". In medical terms, this means someone is tested positive for a disease, but actually does not have it. In our dataset, this means a student was predicted to get **College_Score** at least 65, but actually did not.

$$\text{False Positive Rate} = \frac{\text{False Positive}}{\text{Actual Negative}} = \frac{\text{False Positive}}{\text{True Negative} + \text{False Positive}}$$

Plugging in the numbers from our model, we get

$$\text{False Positive Rate} = \frac{35}{35 + 61} \approx 36.46\%.$$

Given that a student did not get **College_Score** at least 65, there is a 36.46% chance that we predicted him/her to achieve this. We would like to give the benefit of the doubt, saying that the student simply did not do well on the first college entrance exam for early admission. It is possible that he/she was able to get into a better school through taking the second college entrance exam for regular admission.

On the other hand, false negative rate is the probability of an actual positive being classified as a negative. In medical terms, this means someone is tested negative for a disease, but actually has the disease. In our dataset, this means a student actually got **College_Score** at least 65, but we predicted him/her as not achieving this.

$$\text{False Negative Rate} = \frac{\text{False Negative}}{\text{Actual Positive}} = \frac{\text{False Negative}}{\text{True Positive} + \text{False Negative}}$$

False negative rate means how likely the model missed actual positive datapoints, so this rate is the opposite of recall.

$$\text{False Negative Rate} = 1 - \text{Recall}$$

Plugging in the numbers from our model, we get

$$\text{False Negative Rate} = \frac{20}{72 + 20} \approx 21.74\%.$$

Given that a student actually got **College_Score** at least 65, there is a 21.74% chance that we did not predict him/her to achieve this. We need to remember that the model is imperfect, so there always exist students who did better in **College_Score** than the model had expected. To put it differently, **HighSchool_PR** is not a full indicator of achieving **College_Score** at least 65 or not. This is an encouraging message to students who did not do well in **HighSchool_PR**, because they still have a chance in **College_Score** to get admitted to a good school.

In our data analysis, false positive rate and false negative rate have about equal importance. But in certain situations, one can be much more important than the other. For instance, false positive rate is an essential measure in the effectiveness of prompting users to re-enter login information to verify identity for social media. The goal of re-entering credentials is to prevent unauthorized logins, but when people get prompted too many times while using their own account, they would get frustrated and leave the website. This leads to significant revenue loss in business.[76] On the contrary, we are more concerned about the false negative rate in medical testing for a rare disease. The goal of medical testing is to identify as many people with the disease as possible, so that these people can receive timely medical treatment. Hence we are more concerned when the test fails to detect a person who actually has the disease.[77]

### 8.2.3 Sensitivity and Specificity

We are going off on a tangent here, because this section is not directly related to the data of high school and college entrance exam scores. But we think it is important to introduce the concepts of **sensitivity** and **specificity** to the readers, since they are also used to describe the overall testing results, especially in a clinical setting.[78]

Sensitivity means that when a patient actually has the disease (actual positive), the medical test is able to sense it and produce a positive result. That is, the medical test is sensitive enough to identify patients who have the disease.

$$\text{Sensitivity} = \frac{\text{True Positive}}{\text{Actual Positive}} = \frac{\text{True Positive}}{\text{True Positive + False Negative}}.$$

Sensitivity is equivalent to the true positive rate (a.k.a. recall), or the opposite of the false negative rate.

$$\text{Sensitivity} = \text{Recall} = 1 - \text{False Negative Rate}.$$

On the other hand, specificity means when a patient does not have the disease (actual negative), the medical test is able to produce a negative result. That is, the medical test is specific enough that it filters out patients who do not have the disease.

$$\text{Specificity} = \frac{\text{True Negative}}{\text{Actual Negative}} = \frac{\text{True Negative}}{\text{True Negative + False Positive}}.$$

Specificity is equivalent to the true negative rate, or the opposite of the false positive rate.

[76] https://fcase.io/a-major-challenge-false-positives/
[77] https://brownmath.com/stat/falsepos.htm
[78] https://www.healthnewsreview.org/toolkit/tips-for-understanding-studies/understanding-medical-tests-sensitivity-specificity-and-positive-predictive-value/

$$\text{Specificity} = 1 - \text{False Positive Rate}.$$

Let's see an example in medical testing.[79] Assume 0.1% of the population have a specific disease. In a population of 500,000 people, 500 people would have the disease. Now we have a medical test that claims to be 99% accurate, which means the test has 99% sensitivity and 99% specificity. Hence the false positive rate and the false negative rate are both 1%.

- For the 500 people who actually have the disease, 495 tested positive and 5 tested negative.
- For the 499,500 people who do not have the disease, 4,995 people tested positive and the remaining 494,505 people tested negative.

**Given that a patient tested positive, how likely does he/she actually have the disease?**

Using the Bayes theorem, we get

$$\begin{aligned} P(\text{Actual Positive}|\text{Test Positive}) &= \frac{P(\text{Actual Positive} \cap \text{Test Positive})}{P(\text{Test Positive})} \\ &= \frac{P(\text{Actual Positive} \cap \text{Test Positive})}{P(\text{Actual Positive} \cap \text{Test Positive}) + P(\text{Actual Negative} \cap \text{Test Positive})} \\ &= \frac{495}{495 + 4995} \approx 9.16\%. \end{aligned}$$

**The patient has a 9.16% of chance of actually having the disease, despite the positive test outcome. The patient does not have the disease for more than 90% of the time.** Since the disease infects only 0.1% of the population, the medical test creates many false positives, i.e., false alarms. Nevertheless, the test is still informative because given a positive test result, the probability of the patient having the disease increases by 91.6 times.

$$\frac{P(\text{Actual Positive}|\text{Test Positive})}{P(\text{Actual Positive})} = \frac{9.16\%}{0.1\%} = 91.6.$$

The statistical calculation tells us that we do not have to be overly concerned about a positive medical test outcome, because the chances are still low for the person to have the disease. However, upon learning the 99% sensitivity and 99% specificity of the test, many doctors seem to associate a positive test with a high probability of having the disease.[80] If any of the readers become a medical doctor in the future, please remember the lesson here and make better treatment decisions for the patients.

## 8.3 Breakdown by High School Entrance Exam Scores

Let's examine the confusion matrices for each group of **HighSchool_PR**: 0-79, 80-89, 90-94, 95-99. We would like to see if the model performance varies across these groups. Readers can refer to Section 6.1 for the details of this categorization. Before going into the analysis, we need to ensure that each group has a sufficiently large number of respondents within the 188 total records.

The group with **HighSchool_PR** 0-79 has the smallest number of respondents, and 25 is a sufficient sample size. The groups with **HighSchool_PR** 80-89 and 90-94 contain 49 and 44 respondents, respectively. The group with **HighSchool_PR** 95-99 includes 70 respondents, which is the largest of the four categories.

```
HS0to79_ind = which(data_corr$HighSchool_PR >=0 & data_corr$HighSchool_PR <= 79)
HS80to89_ind = which(data_corr$HighSchool_PR >= 80 & data_corr$HighSchool_PR <= 89)
HS90to94_ind = which(data_corr$HighSchool_PR >= 90 & data_corr$HighSchool_PR <= 94)
```

[79] https://math.hmc.edu/funfacts/medical-tests-and-bayes-theorem/

[80] https://www.washingtonpost.com/news/posteverything/wp/2018/10/05/feature/doctors-are-surprisingly-bad-at-reading-lab-results-its-putting-us-all-at-risk/

```r
HS95to99_ind = which(data_corr$HighSchool_PR >= 95 & data_corr$HighSchool_PR <= 99)

print(paste("HighSchool_PR 0-79:",length(HS0to79_ind), "respondents"))
```

```
## [1] "HighSchool_PR 0-79: 25 respondents"
```

```r
print(paste("HighSchool_PR 80-89:",length(HS80to89_ind), "respondents"))
```

```
## [1] "HighSchool_PR 80-89: 49 respondents"
```

```r
print(paste("HighSchool_PR 90-94:",length(HS90to94_ind), "respondents"))
```

```
## [1] "HighSchool_PR 90-94: 44 respondents"
```

```r
print(paste("HighSchool_PR 95-99:",length(HS95to99_ind), "respondents"))
```

```
## [1] "HighSchool_PR 95-99: 70 respondents"
```

Now we can produce a confusion matrix for each of the four groups. Except for **HighSchool_PR** 90-94, all the other three groups contain only one value in the predicted outcomes. The higher **HighSchool_PR** the student had, the higher probability the model would predict him/her to achieve **College_Score** at least 65. This is not completely unexpected, but we would like to emphasize that the model is imperfect. There are still some students with a low **HighSchool_PR** but an impressively good **College_Score**.

- **HighSchool_PR** 0-79 has only FALSE predicted outcomes in **College_Score** at least 65.
- **HighSchool_PR** 80-89 has only FALSE predicted outcomes in **College_Score** at least 65.
- **HighSchool_PR** 90-94 has TRUE and FALSE predicted outcomes in **College_Score** at least 65.
- **HighSchool_PR** 95-99 has only TRUE predicted outcomes in **College_Score** at least 65.

When the predicted outcomes do not include both TRUE and FALSE, the confusion matrix produced in `R` would be 2x1 instead of the full 2x2. Section 8.3.1 shows the incorrect 2x1 confusion matrices, and we will add the missing second column back in Section 8.3.2. Finally, we will interpret these results in Section 8.3.3.

### 8.3.1 Confusion Matrices (Incorrect)

Let's write a function to summarize the actual outcomes and the predicted outcomes into a confusion matrix, using the default settings in `table`. As the readers can see, `table` outputs only the nonzero columns, and that's why we have several 2x1 confusion matrices here.

```r
# Data
actual_65up = data_corr$CS_65up
# Predicted results
predicted_65up = model$fitted.values >= 0.5

confusion_original <- function(HS_inds, actual, predicted) {
  actual = actual[HS_inds]
  predicted = predicted[HS_inds]
  confusion = table(actual, predicted)
  return(confusion)
}
```

Confusion matrix for **HighSchool_PR** 0-79

```r
confusion_0to79 = confusion_original(HS0to79_ind, actual_65up, predicted_65up)
confusion_0to79
```

```
##        predicted
## actual  FALSE
##   FALSE    21
```

```
##   TRUE      4
```

Confusion matrix for **HighSchool_PR** 80-89

```
confusion_80to89 = confusion_original(HS80to89_ind, actual_65up, predicted_65up)
confusion_80to89
```

```
##        predicted
## actual  FALSE
##   FALSE    35
##   TRUE     14
```

Confusion matrix for **HighSchool_PR** 90-94

```
confusion_90to94 = confusion_original(HS90to94_ind, actual_65up, predicted_65up)
confusion_90to94
```

```
##        predicted
## actual  FALSE TRUE
##   FALSE     5   22
##   TRUE      2   15
```

Confusion matrix for **HighSchool_PR** 95-99

```
confusion_95to99 = confusion_original(HS95to99_ind, actual_65up, predicted_65up)
confusion_95to99
```

```
##        predicted
## actual  TRUE
##   FALSE   13
##   TRUE    57
```

### 8.3.2 Confusion Matrices (Correct)

When the confusion matrix has nonzero values in all four cells, we can output the 2x2 confusion matrix. But before doing so, we need to revert the order of FALSE and TRUE in the rows and columns, since `R` sorts names in alphabetical order by default. This can be done by producing the second index (TRUE) before the first index (FALSE).

When all predicted values are FALSE, the confusion matrix is missing a TRUE column and we need to add it back. Similarly, When all predicted values are TRUE, the confusion matrix is missing a FALSE column and we also need to add it back. Finally, we revert the order of FALSE and TRUE in the rows and columns, as in the case of an 2x2 confusion matrix.

By checking the dimensions of the confusing matrices and modifying if necessary, we obtain all four 2x2 confusion matrices for the four categories of **HighSchool_PR**.

```
confusion_subset <- function(HS_inds, actual, predicted) {
  actual = actual[HS_inds]
  predicted = predicted[HS_inds]
  confusion = table(actual, predicted)

  # When the confusion matrix has nonzero values in all four cells
  if ((dim(confusion)[1] == 2) && (dim(confusion)[2] == 2)) {
    # Revert the order of FALSE and TRUE
    confusion = confusion[2:1, 2:1]
    # Exit the function because the operation is complete
    return(confusion)
  }
```

```r
  # When all predicted values are FALSE
  else if (colnames(confusion) == c("FALSE")) {
    confusion = as.table(cbind(confusion,c(0,0)))
    colnames(confusion) = c("FALSE","TRUE")
    names(dimnames(confusion)) = c("actual","predicted")
  }

  # When all predicted values are TRUE
  else if (colnames(confusion) == c("TRUE")) {
    confusion = as.table(cbind(c(0,0), confusion))
    colnames(confusion) = c("FALSE","TRUE")
    names(dimnames(confusion)) = c("actual","predicted")
  }

  # Revert the order of FALSE and TRUE
  confusion = confusion[2:1, 2:1]
  return(confusion)
}
```

Confusion matrix for **HighSchool_PR** 0-79

```r
confusion_0to79 = confusion_subset(HS0to79_ind, actual_65up, predicted_65up)
confusion_0to79
```

```
##        predicted
## actual  TRUE FALSE
##   TRUE     0     4
##   FALSE    0    21
```

Confusion matrix for **HighSchool_PR** 80-89

```r
confusion_80to89 = confusion_subset(HS80to89_ind, actual_65up, predicted_65up)
confusion_80to89
```

```
##        predicted
## actual  TRUE FALSE
##   TRUE     0    14
##   FALSE    0    35
```

Confusion matrix for **HighSchool_PR** 90-94

```r
confusion_90to94 = confusion_subset(HS90to94_ind, actual_65up, predicted_65up)
confusion_90to94
```

```
##        predicted
## actual  TRUE FALSE
##   TRUE    15     2
##   FALSE   22     5
```

Confusion matrix for **HighSchool_PR** 95-99

```r
confusion_95to99 = confusion_subset(HS95to99_ind, actual_65up, predicted_65up)
confusion_95to99
```

```
##        predicted
## actual  TRUE FALSE
##   TRUE    57     0
##   FALSE   13     0
```

### 8.3.3 Interpretations

Table 1 shows the model results for each **HighSchool_PR** category (0-79, 80-89, 90-94, 95-99) – in terms of accuracy, precision, recall, FPR, and FNR. Any metric that cannot be calculated is marked as N/A. The results are quite extreme because the model predicted negative for all **HighSchool_PR** 0-79 and 80-89, and it predicted positive for all **HighSchool_PR** 95-99. Here, positive means achieving **College_Score** 65 or higher, and negative means not achieving this.

The overall accuracy is slightly above 70%, while the accuracy for **HighSchool_PR** 90-94 is below 50%. Nevertheless, the group of **HighSchool_PR** 90-94 is the only category with nonzero values in all four cells of the confusion matrix. All the other three **HighSchool_PR** categories have good accuracy, but their FPR and FNR are either 0% or 100% because all predictions within each group have the same value.

| **HighSchool_PR** | Accuracy | Precision | Recall | FPR | FNR |
|---|---|---|---|---|---|
| 0-79 | 84.00% | N/A | 0% | 0% | 100% |
| 80-89 | 71.43% | N/A | 0% | 0% | 100% |
| 90-94 | 45.45% | 40.54% | 88.23% | 81.48% | 11.76% |
| 95-99 | 81.43% | 81.43% | 100% | 100% | 0% |
| Overall | 70.74% | 67.29% | 78.26% | 36.46% | 21.74% |

Table 1: Model results for each **HighSchool_PR** category

For **HighSchool_PR** 0-79 and 80-89, the precision is not available (N/A) due to division-by-zero. The model made zero positive predictions, i.e., it predicted all students in the two categories not to achieve **College_Score** 65 or higher. Note that these students had a non-zero probability to obtain this achievement, but their predicted probability is less than the threshold of 50%, and that's why the datapoints are predicted as negative. The remaining metrics of **HighSchool_PR** 0-79 and 80-89 are straightforward, because they are either 0% or 100%. The recall is 0% because none of the students in these categories received a positive prediction. The FPR is also 0% because all students who did not achieve **College_Score** 65 or higher were not predicted to do so anyway. The FNR is 100% because all students who actually obtained **College_Score** 65 or higher were not predicted to make it, so these datapoints are regarded as unexpected successes. Finally, the accuracy here equals to the percentage of students who received **College_Score** below 65, and it is reasonable to see a lower accuracy (non-success rate) for **HighSchool_PR** 80-89 than 70-79.

**HighSchool_PR** 90-94 is the most interesting group, because it is the only category where the model predicted both TRUE and FALSE values. Although the accuracy is less than 50%, the recall is impressively high (over 80%). The model predicted 37 out of 44 students in this category to achieve **College_Score** 65 or higher, but 22 of the 37 students did not make it. This resulted in a high FPR and a low FNR. Section 6.2 shows that a **College_Score** of 65 is around the 95th percentile of all college entrance exam takers in Taiwan each year. Therefore, it is reasonable to predict that most students with **HighSchool_PR** 90-94 are going to achieve **College_Score** at least 65. Recall that not all high school students are willing and able to attending college. However, students with **HighSchool_PR** 90-94 do not have the same level of access to academic resources. **HighSchool_PR** of this range can mean attending a top high school in the rural areas.[81] But in Taipei, this **HighSchool_PR** is nowhere sufficient to get admitted into any prestigious high schools (a.k.a. "the top three") within the city.[82] Since we do not have the geographic information of the respondents, we cannot make any inferences on which ones with **HighSchool_PR** 90-94 are more likely to achieve **College_Score** at least 65.

The **HighSchool_PR** 95-99 group contains respondents with top 5% scores in the high school entrance exam, and that's why the model predicted all of them to achieve **College_Score** at least 65. The recall is 100% due to all datapoints being predicted as TRUE. The FPR is also 100% because all respondents in this category who did not achieve **College_Score** at least 65 were predicted to do so. The FNR is 0% because none of the respondents were predicted not to achieve this. The accuracy is over 80% because most respondents with **HighSchool_PR** 95-99 actually obtained **College_Score** at least 65. The 4x4 table

[81] https://bit.ly/3qxuqMw
[82] https://cclccl-life.blogspot.com/2013/06/blog-post_9.html

in Section 6.3 shows that 23 of them (32.86%) got **College_Score** between 65 and 69, while 34 of them (48.57%) got **College_Score** between 70 and 75. One extension would be increasing the **College_Score** cutoff point to 70 for this group, but we do not think this is practical without external data about these respondents.

In our opinion, **HighSchool_PR** 95-99 does little in distinguishing students' academic capabilities within the group, because it is possible to drop one percentile by getting just one critical question wrong on the high school entrance exam. Prior to 2009, there was a huge score difference between full marks and missing by just one question.[83] This could result in a significant difference of **HighSchool_PR**, because missing five questions in a single subject (full marks on the other four) would have a better score than missing one question in each of the five subjects. Hence it is not meaningful to evaluate students' academic performance solely from **HighSchool_PR**. Similar to the case in **HighSchool_PR** 90-94, we cannot make inferences about the within-group differences of **HighSchool_PR** 95-99. In rural areas, students with **HighSchool_PR** 95-99 already meet the admission requirements for the top high school in their region. Almost all of them would attend that school unless they move to a different region. As a comparison, Taipei's first choice high school requires **HighSchool_PR** at least 99, second choice requires at least 98, and third choice requires at least 97.[84] As a result, students in Taipei with **HighSchool_PR** 95 or 96 often end up attending local high schools outside the top tier. These schools are still high-quality and provide ample resources, but they are not as prestigious as the top three choices.

# 9 Model Validation: Out-of-Sample Prediction

The goal of building this model is to optimize the performance for future input, i.e., incoming students who just obtained their **HighSchool_PR** scores. We need to use the model to predict new data, and this validation method is called out-of-sample prediction. That is, the model has to be able to predict data outside the training sample.

In-sample prediction (Chapter 8) is insufficient because we need to test on unseen data to **avoid overfitting**.[85] But why is overfitting bad? Because the model would do well on the existing data but perform poorly on the new data, which is undesirable. This is similar to a student who memorizes the answers to score 100% on quizzes without understanding the actual content. Then this student may not do well on the final exam because he/she has not seen the questions before. In order to measure the student's grasp of the knowledge, the instructor usually gives exam questions similar to the practice questions, but not exactly the same.

Some readers may be wondering how to get "new" data to perform out-of-sample prediction, and the good news is that we already have them. New data means **previously unseen** data by the model; in other words, the data was not involved in training the model. Although training the model requires data, we do not have to feed in all 188 records at once. We can use a part of the records to train the model, and leverage the remaining data to test the model for performance evaluation. In this way, the latter part of the data are considered "new" because they are not seen by the model beforehand. The data involved in the training phase is called the training dataset, and the data used for testing is called the testing dataset.

In this chapter, we demonstrate two methods to implement out-of-sample prediction. The first method is using **separate training and testing datasets** to validate the model, where the two datasets are mutually exclusive. We train the model on the training set, and test the model on the testing set. The second method is **cross validation**, which involves partitioning the data into a number of subsets, then we reserve one subset for testing and train the model on all the remaining subsets. Each subset take turns to be used for testing, and finally we combine the results to estimate the overall prediction performance.

## 9.1 Separate Training and Testing Datasets

In this section, **randomly divide the data into a training set and a testing set**. Then we use the training set to train the model for the parameters, and use the testing set to evaluate the model performance.

[83] https://news.ltn.com.tw/news/life/paper/217325
[84] https://chendaneyl.pixnet.net/blog/post/31436728
[85] https://elitedatascience.com/overfitting-in-machine-learning

In other words, the model is trained on some data independent of the testing set, because it would not see the testing data beforehand. Using unseen data is helpful to get a better measure of the prediction power of the model. An extension is to divide the data into **training, validation, and testing** sets. We still use the training set to train the model, but we use the validation set to fine-tune the model parameters, i.e., hyperparameter tuning.[86] Finally, we evaluate the model using the testing set. By incorporating a validation set as a mid-step, we do not look at the model results for the testing set until the end. This further reduces the risk of overfitting the testing data. But since our logistic regression model does not involve hyperparameter tuning, we use only the training and testing sets for simplicity.

### 9.1.1 Implementation

Below is the code to divide the data into the training and testing partitions. We randomly selected 50% of the data (94 out of the 188 records) to be in the training part, and the remaining 50% are in the testing part. This can be done by a random permutation of the indices 1-194. Then the first half of the indices correspond to the training records, and the second half of the indices correspond to the testing records. We set a random seed to ensure reproducibility.

```
set.seed(10)

nn = nrow(data_corr) # total 188 rows of data

row_inds = c(1:nn)

ind_permute = sample(row_inds)

train_inds = ind_permute[1:94]
test_inds = ind_permute[95:188]
```

The `train_inds` are the indices for the training part of the data:

```
print(train_inds)
```

```
##  [1] 137  74 112 183  72 182 167  88  15 143 170 187 162  24  13  95 136 110   7
## [20] 155  86  82  29 166 121  92  50 109 154 101 122  33 135 160  68  93 114 181
## [39]  51  32  11  79 163  91  42  78 174 105 117  26  89  48 180 171 175  61 132
## [58]  14  35  10 177  58  39  16 172  31 129 159 150  53  63 164 161  47 126 120
## [77] 138  18   3 168 144 158  64  59 147  77  90 179 146  34 106   4 118  20
```

The `test_inds` are the indices for the testing part of the data:

```
print(test_inds)
```

```
##  [1]  96 153  87 188 173  54  57  27   9  80 145 116 104 108  73 184  25 130 141
## [20]  46 113  22 142   6  40  71 119 149 123 134 169 176  45  23  17  43 140   1
## [39]  97  30 165 131 115  65  62  38 102 152  55   2 186  28  76  12  52  81  75
## [58] 100 156  37  98  49 103 107  67   8  69   5  83 111 124 125  60  99 157  70
## [77] 185  36  41  21  85 133  44 127  19  94  84 148 151 178 139  56  66 128
```

We can sort each set of the indices in ascending order, so it will be easier to refer to them later, i.e., better readability. When we obtain the testing results, the records would be in the same order as in the original dataset.

```
train_inds = sort(train_inds)
test_inds = sort(test_inds)
```

After sorting, the 94 training indices are in ascending order.

[86] https://towardsdatascience.com/train-validation-and-test-sets-72cb40cba9e7

```r
print(train_inds)
```

```
##  [1]   3   4   7  10  11  13  14  15  16  18  20  24  26  29  31  32  33  34  35
## [20]  39  42  47  48  50  51  53  58  59  61  63  64  68  72  74  77  78  79  82
## [39]  86  88  89  90  91  92  93  95 101 105 106 109 110 112 114 117 118 120 121
## [58] 122 126 129 132 135 136 137 138 143 144 146 147 150 154 155 158 159 160 161
## [77] 162 163 164 166 167 168 170 171 172 174 175 177 179 180 181 182 183 187
```

The 94 testing indices are also sorted in ascending order.

```r
print(test_inds)
```

```
##  [1]   1   2   5   6   8   9  12  17  19  21  22  23  25  27  28  30  36  37  38
## [20]  40  41  43  44  45  46  49  52  54  55  56  57  60  62  65  66  67  69  70
## [39]  71  73  75  76  80  81  83  84  85  87  94  96  97  98  99 100 102 103 104
## [58] 107 108 111 113 115 116 119 123 124 125 127 128 130 131 133 134 139 140 141
## [77] 142 145 148 149 151 152 153 156 157 165 169 173 176 178 184 185 186 188
```

Now we slice the data into the training and testing parts using the two sets of indices.

```r
train_data = data_corr[train_inds,]
test_data = data_corr[test_inds,]
```

Then we train the logistic regression model using the 188 records in the training part, and the model summary shows the coefficient point estimates along with the standard error.

```r
train_model = glm(CS_65up ~ HighSchool_PR, data=train_data, family="binomial")
summary(train_model)
```

```
##
## Call:
## glm(formula = CS_65up ~ HighSchool_PR, family = "binomial", data = train_data)
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -15.2067     3.7687  -4.035 5.46e-05 ***
## HighSchool_PR   0.1661     0.0408   4.072 4.66e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 130.27  on 93  degrees of freedom
## Residual deviance: 104.80  on 92  degrees of freedom
## AIC: 108.8
##
## Number of Fisher Scoring iterations: 5
```

Next, we use the trained model to predict the testing part of the data. The function `predict.glm` allows us to fit the generalized linear model (GLM) on new data. The type 'response' gives the predicted probabilities. The output is a numeric vector with the predicted probabilities, and the header is the record index from the original data. For example, the 1st record in the original data is included in the testing part, and the model predicts the respondent to have a 0.5% probability of obtaining **College_Score** 65 or higher.

```r
test_prob = predict.glm(train_model, test_data, type="response")
round(test_prob, digits=3)
```

```
##     1     2     5     7     9    10    13    18    21    23    24    25    27
```

```
## 0.005 0.005 0.320 0.023 0.640 0.357 0.640 0.745 0.052 0.222 0.561 0.745 0.677
##    29    30    32    38    39    40    42    43    45    46    47    48    51
## 0.222 0.519 0.640 0.396 0.776 0.640 0.222 0.776 0.561 0.601 0.713 0.776 0.677
##    54    56    57    58    59    62    64    67    68    69    72    73    74
## 0.437 0.396 0.519 0.170 0.677 0.070 0.032 0.561 0.148 0.776 0.111 0.060 0.095
##    76    78    79    83    84    87    89    90    92   100   102   103   104
## 0.745 0.745 0.148 0.713 0.601 0.519 0.111 0.478 0.745 0.128 0.478 0.561 0.601
##   105   106   108   109   110   113   114   117   119   121   122   125   129
## 0.776 0.320 0.396 0.776 0.601 0.111 0.561 0.437 0.252 0.776 0.320 0.713 0.320
##   130   131   134   135   137   138   140   141   146   147   148   149   152
## 0.601 0.252 0.776 0.357 0.252 0.776 0.561 0.745 0.745 0.128 0.478 0.437 0.776
##   155   156   158   159   160   163   164   172   176   180   184   186   192
## 0.713 0.776 0.745 0.437 0.713 0.195 0.285 0.037 0.357 0.745 0.001 0.396 0.601
##   193   194   197
## 0.745 0.052 0.027
```

Then we follow the procedures in Section 8.1 to convert the test probabilities into binary classification results, i.e., the confusion matrix.

```r
# Convert the test probabilities into binary classification results
test_actual_65up = test_data$CS_65up
test_pred_65up = test_prob > 0.5

# Confusion matrix
test_confusion = table(test_actual_65up, test_pred_65up)
# revert the order of FALSE and TRUE
test_confusion = test_confusion[2:1, 2:1]

test_confusion
```

```
##                 test_pred_65up
## test_actual_65up TRUE FALSE
##            TRUE    34    12
##            FALSE   14    34
```

We can also produce the percentage version of the confusion matrix.

```r
prop.table(test_confusion)
```

```
##                 test_pred_65up
## test_actual_65up      TRUE     FALSE
##            TRUE  0.3617021 0.1276596
##            FALSE 0.1489362 0.3617021
```

Now we show the number of true positives, false negatives, false positives, and false negatives.

```r
# row = actual_65up, column = predicted_65up
tp = test_confusion["TRUE","TRUE"]
fn = test_confusion["TRUE","FALSE"]
fp = test_confusion["FALSE","TRUE"]
tn = test_confusion["FALSE","FALSE"]

print(paste(tp,fn,fp,tn))
```

```
## [1] "34 12 14 34"
```

We can also calculate the evaluation metrics for the predictive model. The process is similar to Section 8.2.

$$
\begin{aligned}
\text{Accuracy} &= \frac{TP+TN}{TP+FN+FP+TN} = \frac{34+34}{34+12+14+34} = \frac{68}{94} \approx 72.34\% \\
\text{Precision} &= \frac{TP}{TP+FP} = \frac{34}{34+14} = \frac{34}{48} \approx 70.83\% \\
\text{Recall} &= \frac{TP}{TP+FN} = \frac{34}{34+12} = \frac{34}{46} \approx 73.91\% \\
\text{False Positive Rate (FPR)} &= \frac{FP}{TN+FP} = \frac{14}{34+14} = \frac{14}{48} \approx 29.17\% \\
\text{False Negative Rate (FNR)} &= \frac{FN}{TP+FN} = \frac{12}{34+12} = \frac{12}{46} \approx 26.09\%
\end{aligned}
$$

We also compare the results with the in-sample prediction, and they show similar trends. The accuracy, precision, and recall all hover around 70%.

| | Accuracy | Precision | Recall | FPR | FNR |
|---|---|---|---|---|---|
| In-Sample Prediction | 70.74% | 67.29% | 78.26% | 36.46% | 21.74% |
| Separate Training and Testing | 72.34% | 70.83% | 73.91% | 29.17% | 26.09% |

Table 2: Comparison of results with in-sample prediction

### 9.1.2 Organizing the Code for Reusability

We have demonstrated how to train the logistic regression model on a part of the data, and test the model on the remaining data. But there is one problem – the code is messy and hence not reusable. If we are going to build another training-testing framework, we may have to copy-paste lots of code from Section 9.1.1, which is undesirable. We should avoid excessive copy-pasting because this is prone to mistakes, making the code more difficult to debug.

A better solution is to incorporate repetitive code into a function, so that we can keep the same work in one place. In software development, "don't repeat yourself" (DRY) is a principle to reduce code repetitions (Foote, 2014). When we change this part of the program, we only need to edit the code within the function. The modifications would automatically be performed anytime the function is called. In this way, the code can be easily reused and maintained.

As a first example, we need to convert the predictive probabilities into the binary classification results, and show them in a confusion matrix. This compound task is performed in almost every model validation involving binary classifications, so we should encapsulate the task into a function. This function compares the test probabilities with their ground truth (0/1), and outputs the number of true positives, false negatives, false positives, and true negatives. We predict a datapoint to be positive if the estimated probability is at or above a given threshold, which is set to 0.5 by default. Otherwise, we predict the datapoint to be negative.

```
prob_to_matrix <- function(test_data, test_prob, threshold=0.5) {
  # Convert the test probabilities into binary classification results.
  # Threshold should be between 0 and 1, set to 0.5 by default.

  test_actual_65up = test_data$CS_65up
  test_pred_65up = test_prob >= threshold

  # Confusion matrix
  test_confusion = table(test_actual_65up, test_pred_65up)
  # revert the order of FALSE and TRUE
  test_confusion = test_confusion[2:1, 2:1]
```

```
  return(test_confusion)
}
```

We can call the `prob_to_matrix` function to obtain the confusion matrix, and the output is the same.

```
another_test = prob_to_matrix(test_data, test_prob)
another_test
```

```
##                 test_pred_65up
## test_actual_65up TRUE FALSE
##            TRUE    34    12
##            FALSE   14    34
```

The results in Table 2 are for separate training and testing sets from a single random seed. We would like to try more versions of such out-of-sample prediction, so we created the function `train_and_test` to automate the procedure. Note that this procedure calls `prob_to_matrix` at the end. We wrote the latter as a single function because we may also use it in other types of model validation. Eventually, we can run this function multiple times and take the average of the accuracy/precision/recall/etc.

```
train_and_test <- function(data, seed) {
  # Automate the procedure of using training and testing datasets
  # for out-of-sample model validation.

  # Input: data_corr, random_seed
  # Output: confusion_matrix

  set.seed(seed)
  nn = nrow(data)
  row_inds = c(1:nn)
  ind_permute = sample(row_inds)
  mid_pt = floor(nn/2) # round down

  # Randomly split the data into 50% training and 50% testing
  train_inds = ind_permute[1:mid_pt]
  test_inds = ind_permute[(mid_pt+1):nn]
  train_inds = sort(train_inds)
  test_inds = sort(test_inds)

  train_data = data[train_inds,]
  test_data = data[test_inds,]

  train_model = glm(CS_65up ~ HighSchool_PR, data=train_data, family="binomial")
  # summary(train_model)

  test_prob = predict.glm(train_model, test_data, type="response")
  # round(test_prob, digits=3)

  test_confusion = prob_to_matrix(test_data, test_prob)

  return(test_confusion)
}
```

With this function, we can reproduce the predictive outcomes using the same random seed.

```
train_and_test(data_corr, seed=10)
```

```
##                 test_pred_65up
```

```
## test_actual_65up TRUE FALSE
##            TRUE    34    12
##            FALSE   14    34
```

We can try a different random seed, and obtain results with a different split of training/testing data.

```
train_and_test(data_corr, seed=123)
```

```
##                 test_pred_65up
## test_actual_65up TRUE FALSE
##            TRUE    38     9
##            FALSE    8    39
```

Let's try five iterations with different random seeds and output the results. We will get five confusion matrices, and we can summarize the numbers across them. The main reason to try multiple iterations is to avoid getting an unlucky draw, i.e., a single partition that leads to extreme outcomes.

In the code, we generate the sequence of random seeds from a sequence of random numbers between 1 and 1000 without replacement. In this way, we only need to set a single random seed to get all five runs (which we can increase in the future).

```
set.seed(37)
runs = 5

# Discrete uniform distribution:
# Generate a sequence of random numbers between 1 and 1000
# (sample without replacement)
seed_each = sample(1:1000, runs, replace=F)

# Initialize the list with size = number of runs.
# Don't start with an empty list and append elements later,
# because the append function may not work for matrix elements.

out_matrices = rep(list("results"), runs)

for (iter in 1:runs) {
  output = train_and_test(data_corr, seed=seed_each[iter])
  out_matrices[[iter]] = output
}

out_matrices
```

```
## [[1]]
##                 test_pred_65up
## test_actual_65up TRUE FALSE
##            TRUE    36    11
##            FALSE   15    32
##
## [[2]]
##                 test_pred_65up
## test_actual_65up TRUE FALSE
##            TRUE    31    14
##            FALSE   12    37
##
## [[3]]
##                 test_pred_65up
## test_actual_65up TRUE FALSE
```

```
##            TRUE    36    11
##            FALSE   23    24
##
## [[4]]
##                 test_pred_65up
## test_actual_65up TRUE FALSE
##            TRUE    38     9
##            FALSE   14    33
##
## [[5]]
##                 test_pred_65up
## test_actual_65up TRUE FALSE
##            TRUE    40    10
##            FALSE   16    28
```

For each confusion matrix, we need to calculate the accuracy, precision, recall, FPR, and FNR. We calculated these by hand earlier, and now it is time to do them in `R` code for reproducibility. We write a function to do the calculations, because these would be reused in other model validation schemes. We also use the first confusion matrix to demonstrate how this function works.

```r
confusion_to_measures <- function(output) {
  tp = output[1,1]
  fn = output[1,2]
  fp = output[2,1]
  tn = output[2,2]

  accuracy = (tp+tn)/(tp+fn+fp+tn)
  precision = tp/(tp+fp)
  recall = tp/(tp+fn)
  fpr = fp/(tn+fp)
  fnr = fn/(tp+fn)

  measures = c(accuracy, precision, recall, fpr, fnr)
  names(measures) = c("Accuracy","Precision","Recall","FPR","FNR")

  return(measures)
}

sample_output = confusion_to_measures(out_matrices[[1]])
round(sample_output, digits = 4)
```

```
##  Accuracy Precision    Recall       FPR       FNR
##    0.7234    0.7059    0.7660    0.3191    0.2340
```

Then we output the metrics of each iteration to a table, using a for loop to run through the iterations. We converted this process into the function `combine_results`, because we need to use the same script for k-fold cross validation as well.

```r
combine_results <- function(out_matrices) {
  # Combine the output results
  # Input: out_matrices (list of matrices)
  # Output: out_measures (matrix array)

  runs = length(out_matrices)

  out_measures = c(0,0,0,0,0,0)
```

```r
  names(out_measures) = c("Iteration","Accuracy","Precision","Recall","FPR","FNR")

  for (iter in 1:runs) {
    output = confusion_to_measures(out_matrices[[iter]])
    measures = c(iter, output)
    out_measures = rbind(out_measures, measures)
  }

  row.names(out_measures) = rep(c(""), runs+1) # remove row names
  out_measures = out_measures[-1,] # remove the first placeholder row

  return(out_measures)
}

out_measures = combine_results(out_matrices)

# out_measures
round(out_measures, digits=4)
```

```
##  Iteration Accuracy Precision Recall    FPR    FNR
##          1   0.7234    0.7059 0.7660 0.3191 0.2340
##          2   0.7234    0.7209 0.6889 0.2449 0.3111
##          3   0.6383    0.6102 0.7660 0.4894 0.2340
##          4   0.7553    0.7308 0.8085 0.2979 0.1915
##          5   0.7234    0.7143 0.8000 0.3636 0.2000
```

We also calculate the average of the five iterations for each metric. The accuracy, precision, and recall all hover around 70% as expected, just like in Table 2 in Section 9.1.1.

```r
calc_average <- function(out_measures) {
  avg_results = c(mean(out_measures[,"Accuracy"]),
            mean(out_measures[,"Precision"]),
            mean(out_measures[,"Recall"]),
            mean(out_measures[,"FPR"]), mean(out_measures[,"FNR"]))

  names(avg_results) = c("Accuracy","Precision","Recall","FPR","FNR")

  return(avg_results)
}

average = calc_average(out_measures)

# average
round(average, digits=4)
```

```
##  Accuracy Precision    Recall       FPR       FNR
##    0.7128    0.6964    0.7659    0.3430    0.2341
```

We are going to have fewer descriptions in the result evaluation of later sections, because we assume that at this point, the readers would already be familiar with the relevant concepts.

## 9.2 Cross Validation

Next, we are going to talk about **cross validation**. The "cross" means that each record in the data has the opportunity to serve as the training set AND the testing set (obviously, not at the same time). Cross validation involves partitioning data into a number of subsets, then we reserve one subset for testing and

train the model on all the remaining subsets. Each subset take turns to be used for testing, and finally we combine the results to estimate the overall prediction performance. Two common cross validation methods are **k-fold cross validation** and **leave-one-out cross validation**. We will demonstrate both ways of cross validation in this section.

### 9.2.1 K-fold Cross Validation

In **k-fold cross validation**, we randomly divide the data into $k$ subsets to cross-validate each other. (Typically $k = 10$.) For each round of validation, we train the model on the $k - 1$ subsets and test the model on the one subset which was excluded in the training. Finally, we combine all $k$ rounds of validation, and each subset gets its predicted result for performance evaluation.

To implement 10-fold cross validation, we divide the data into 10 partitions of near-equal sizes. We have 188 records, so there should be eight partitions of 19 records and two partitions of 18 records. We store the indices of each partition in the `partition_list`.

```
# 10-fold cross validation:
# Divide 188 records into 10 partitions of near-equal size

# Number of records in each partition:
# 19, 19, 19, 19, 19, 19, 19, 19, 18, 18
k_fold = c(19, 19, 19, 19, 19, 19, 19, 19, 18, 18)
k_accumulate = c(19, 38, 57, 76, 95, 114, 133, 152, 170, 188)

k_fold_partition <- function(data, k_fold, seed) {
  # Generate the 10 partitions for the data

  set.seed(seed)
  nn = nrow(data) # total 188 rows of data
  row_inds = c(1:nn)
  ind_permute = sample(row_inds)
  # random permutation of row indices
  # => prepare for the training/testing partitions

  partition_list = list(0,0,0,0,0,0,0,0,0,0)

  # Need to sort the indices within each partition
  partition_list[[1]] = sort(ind_permute[1:k_fold[1]])
  for (ii in 2:length(k_fold)) {
    start = k_accumulate[ii-1]+1
    end = start + k_fold[ii] - 1
    partition_list[[ii]] = sort(ind_permute[start:end])
  }

  return(partition_list)
}

partition_list = k_fold_partition(data_corr, k_fold, seed=21)
partition_list
```

```
## [[1]]
##  [1]   3  16  21  47  57  62  63  67  94 106 107 115 123 138 142 157 161 166 170
##
## [[2]]
##  [1]   1   5  33  49  71  72  74  91  92  96  98 105 112 124 131 137 151 159 184
##
```

```
## [[3]]
##  [1]   8  11  15  20  40  45  50  56  65  66  86 117 125 126 135 145 160 162 165
##
## [[4]]
##  [1]   6  22  34  37  53  58  69  89 101 103 104 116 127 132 144 146 172 174 175
##
## [[5]]
##  [1]  28  29  31  35  55  80  81  88 111 113 118 140 152 154 163 173 176 186 187
##
## [[6]]
##  [1]  25  36  44  51  52  59  61  73 110 121 128 129 130 139 150 153 167 177 185
##
## [[7]]
##  [1]   4  10  12  18  19  27  38  43  46  68  75  84  85  87 136 148 156 164 168
##
## [[8]]
##  [1]   7  23  42  64  77  90  93 100 108 109 119 134 143 147 158 178 179 181 183
##
## [[9]]
##  [1]   9  14  26  30  39  41  54  76  78  79  82  95  99 114 120 141 155 180
##
## [[10]]
##  [1]   2  13  17  24  32  48  60  70  83  97 102 122 133 149 169 171 182 188
```

After obtaining the 10 partitions, we reserve one partition as the testing set and feed the other 9 partitions into the training. Each partition gets the chance to be the testing set, and we obtain all 10 confusion matrices.

Other `R` packages may have functions to handle k-fold cross validation, but we decided to write our own code to show the readers how the method is implemented from scratch. Moreover, this allows maximum flexibility for us to modify the code for future needs.

```r
k_fold_train_test <- function(data, partition_list, k_fold) {
  # Training and testing process for k-fold cross validation

  # Use the partitions for training and testing
  partition_probs = list(0,0,0,0,0,0,0,0,0,0)
  partition_matrices = list(0,0,0,0,0,0,0,0,0,0)

  for (exclude in 1:length(k_fold)) {
    # Testing parts
    testing_with_k = partition_list[[exclude]]
    test_kfold_data = data_corr[testing_with_k,]

    # Training parts
    # partition_list[-exclude] shows all elements except the exclude.
    training_without_k = unlist(partition_list[-exclude])
    # integer vector of training indices
    train_kfold_data = data_corr[training_without_k,]

    train_kfold_model = glm(CS_65up ~ HighSchool_PR,
                            data=train_kfold_data, family="binomial")
    # summary(train_kfold_model)
    test_kfold_prob = predict.glm(train_kfold_model,
                                  test_kfold_data, type="response")
    # type="response" gives the predicted probabilities
```

```r
    # Store the predicted probabilities of each partition in a list
    partition_probs[[exclude]] = test_kfold_prob

    # Store the confusion matrix of each partition in another list
    partition_matrices[[exclude]] = prob_to_matrix(test_kfold_data, test_kfold_prob)
  }

  # partition_probs
  return(partition_matrices)
}

partition_matrices = k_fold_train_test(data_corr, partition_list, k_fold)
partition_matrices
```

```
## [[1]]
##                 test_pred_65up
## test_actual_65up TRUE FALSE
##            TRUE     7     3
##            FALSE    1     8
##
## [[2]]
##                 test_pred_65up
## test_actual_65up TRUE FALSE
##            TRUE     7     1
##            FALSE    4     7
##
## [[3]]
##                 test_pred_65up
## test_actual_65up TRUE FALSE
##            TRUE     7     1
##            FALSE    1    10
##
## [[4]]
##                 test_pred_65up
## test_actual_65up TRUE FALSE
##            TRUE    10     1
##            FALSE    3     5
##
## [[5]]
##                 test_pred_65up
## test_actual_65up TRUE FALSE
##            TRUE     4     2
##            FALSE    5     8
##
## [[6]]
##                 test_pred_65up
## test_actual_65up TRUE FALSE
##            TRUE     6     5
##            FALSE    4     4
##
## [[7]]
##                 test_pred_65up
## test_actual_65up TRUE FALSE
```

```
##            TRUE     8     2
##            FALSE    3     6
##
## [[8]]
##                 test_pred_65up
## test_actual_65up TRUE FALSE
##            TRUE     9     2
##            FALSE    5     3
##
## [[9]]
##                 test_pred_65up
## test_actual_65up TRUE FALSE
##            TRUE     9     2
##            FALSE    2     5
##
## [[10]]
##                 test_pred_65up
## test_actual_65up TRUE FALSE
##            TRUE     5     1
##            FALSE    3     9
```

Now we summarize the results in k-fold cross-validation, i.e., combine all 10 confusion matrices into one. We write a function to encapsulate the sum-up process of the confusion matrices, in order to reuse the code later.

```r
sum_up_confusion <- function(k_fold, partition_matrices) {
  # Sum up all k confusion matrices into a single one.
  tp = 0
  fp = 0
  fn = 0
  tn = 0

  for (part in 1:length(k_fold)) {
    tp = tp + partition_matrices[[part]][1]
    fp = fp + partition_matrices[[part]][2]
    fn = fn + partition_matrices[[part]][3]
    tn = tn + partition_matrices[[part]][4]
  }

  # Use an existing confusion matrix as a template.
  k_fold_table = partition_matrices[[1]]

  k_fold_table[1,1] = tp
  k_fold_table[1,2] = fn
  k_fold_table[2,1] = fp
  k_fold_table[2,2] = tn

  return(k_fold_table)
}

k_fold_table = sum_up_confusion(k_fold, partition_matrices)
k_fold_table
```

```
##                 test_pred_65up
## test_actual_65up TRUE FALSE
##            TRUE    72    20
```

```
##            FALSE   31    65
```

After obtaining the 10-fold cross validation results in a single confusion matrix, we can calculate the accuracy, precision, recall, FPR, FNR using the `confusion_to_measures` function from the previous section.

```
k_fold_results = confusion_to_measures(k_fold_table)
round(k_fold_results, digits=4)
```

```
##  Accuracy Precision    Recall       FPR       FNR
##    0.7287    0.6990    0.7826    0.3229    0.2174
```

Since 10-fold cross validation involves randomly partitioning the data into 10 parts of nearly equal size, we can try different random seeds to see how the results change. For each of the five random seeds, the confusion matrices are quite similar.

```
set.seed(37)
runs = 5
# Discrete uniform distribution:
# Generate a sequence of random numbers between 1 and 1000
# (sample without replacement)
seed_each = sample(1:1000, runs, replace=F)

for (iter in 1:runs){
  partition_list = k_fold_partition(data_corr, k_fold,seed=seed_each[iter])
  partition_matrices = k_fold_train_test(data_corr, partition_list, k_fold)
  out_matrices[[iter]] = sum_up_confusion(k_fold, partition_matrices)
  print(out_matrices[[iter]])
}
```

```
##                 test_pred_65up
## test_actual_65up TRUE FALSE
##            TRUE    72    20
##            FALSE   35    61
##                 test_pred_65up
## test_actual_65up TRUE FALSE
##            TRUE    72    20
##            FALSE   34    62
##                 test_pred_65up
## test_actual_65up TRUE FALSE
##            TRUE    72    20
##            FALSE   33    63
##                 test_pred_65up
## test_actual_65up TRUE FALSE
##            TRUE    72    20
##            FALSE   33    63
##                 test_pred_65up
## test_actual_65up TRUE FALSE
##            TRUE    72    20
##            FALSE   32    64
```

We also output the results of the five iterations.

```
out_measures = combine_results(out_matrices)

# out_measures
round(out_measures, digits=4)
```

```
##  Iteration Accuracy Precision Recall    FPR    FNR
```

```
##          1   0.7074    0.6729 0.7826 0.3646 0.2174
##          2   0.7128    0.6792 0.7826 0.3542 0.2174
##          3   0.7181    0.6857 0.7826 0.3438 0.2174
##          4   0.7181    0.6857 0.7826 0.3438 0.2174
##          5   0.7234    0.6923 0.7826 0.3333 0.2174
```

Then we calculate the average of the five iterations, and the mean accuracy is slightly over 70%. Note that the actual numbers can vary depending on the random seed.

```
average = calc_average(out_measures)

# average
round(average, digits=4)
```

```
##  Accuracy Precision    Recall       FPR       FNR
##    0.7160    0.6832    0.7826    0.3479    0.2174
```

### 9.2.2 Leave-one-out Cross Validation

In **leave-one-out cross validation**, each record is considered an independent subset. This is essentially setting $K$ to be the number of total records in the data, say $N$. We train the model on the $N-1$ records and test the model on the one left-out record. This allows each record to get its own prediction. The key advantage is that the results are a relatively accurate estimate of the model performance.[87] Note that when $N$ is extremely large, the computational cost would be high because we need to perform $N$ rounds of validation with $N-1$ records each. The complexity is $O(N^2)$.

Here is the code for leave-one-out cross validation. Since each record is predicted by all the other records in the data, no randomness is involved in creating the data split. Hence we do not have to set a random seed in the process.

```
nn = nrow(data_corr) # total 188 rows of data

prob_leave1out = rep(c(-1), nn)

for (ii in 1:nn) {
  data_test = data_corr[ii,] # reserve one record for testing
  data_exclude = data_corr[-ii,]

  train_leave1out = glm(CS_65up ~ HighSchool_PR, data=data_exclude, family="binomial")
  # summary(train_leave1out)

  test_leave1out = predict.glm(train_leave1out, data_test, type="response")
  # type="response" gives the predicted probabilities

  # Store the predicted probability to the general list
  prob_leave1out[ii] = test_leave1out
}
```

Now we summarize the predictive probabilities in a single confusion matrix. The results are exactly the same as the in-sample prediction in Section 8, which may be a coincidence.

```
matrix_leave1out = prob_to_matrix(data_corr, prob_leave1out)
matrix_leave1out
```

```
##                 test_pred_65up
## test_actual_65up TRUE FALSE
```

[87] https://machinelearningmastery.com/loocv-for-evaluating-machine-learning-algorithms/

```
##            TRUE    72    20
##            FALSE   35    61
```

We also calculate the five metrics: accuracy, precision, recall, FPR, and FNR.

```
leave1out_results = confusion_to_measures(matrix_leave1out)
round(leave1out_results, digits=4)
```

```
##  Accuracy Precision    Recall       FPR       FNR
##    0.7074    0.6729    0.7826    0.3646    0.2174
```

## 9.3 Comparison of Results

Table 3 summarizes the results of in-sample prediction and out-of-sample prediction. Separate training & testing and k-fold cross validation are the average results of five iterations each. The accuracy is slightly above 70%, the precision is around 68%, and the recall is approximately 78%. Since the outcomes are similar across each method, we are not concerned about overfitting in the logistic regression model. Note that the logistic regression is straightforward to run and does not require hyperparameter tuning.

**Remark**: In contrast, some machine learning models (e.g. decision trees) have a large number of parameters, and they are at more risk of overfitting when we tune the parameters to get a lower prediction error. Hyperparameter tuning is also necessary for ridge regression and Lasso regression, both of which contain a penalty term to regulate the number of coefficients in the model.

| | Accuracy | Precision | Recall | FPR | FNR |
|---|---|---|---|---|---|
| In-Sample Prediction | 70.74% | 67.29% | 78.26% | 36.46% | 21.74% |
| Separate Training & Testing (Average) | 71.28% | 69.64% | 76.59% | 34.30% | 23.41% |
| K-Fold Cross Validation (Average) | 71.60% | 68.32% | 78.26% | 34.79% | 21.74% |
| Leave-one-out Cross Validation | 70.74% | 67.29% | 78.26% | 36.46% | 21.74% |

Table 3: Comparison of results with in-sample and out-of-sample prediction

This model uses **HighSchool_PR** scores to predict whether a student would get **College_Score** at least 65 or not. But the model is imperfect in prediction, just as we explained in Section 8.2.

- The **precision** is the number of true positives divided by the predicted positives. This means among the students with good **HighSchool_PR** scores, around 68% of them achieved **College_Score** at least 65 three years later.[88]
- The **recall** is the number of true positives divided by the actual positives. This means among the students with **College_Score** at least 65, approximately 78% them had good **HighSchool_PR** scores three years ago.
- The **FPR (false positive rate)** is the number of false positives divided by the actual negatives. This means among the students with **College_Score** 64 or below, about 35% of them were originally predicted to have **College_Score** at least 65.
- The **FNR (false negative rate)** is the number of false negatives divided by the actual positives. This means among the students with **College_Score** at least 65, slightly over 20% of them were "pleasant surprises" because we did not predict them to achieve such scores given their **HighSchool_PR**.

In summary, given the student's **HighSchool_PR**, the model is only about 70% accurate to predict **College_Score** at least 65 or not. In practical terms, if a student obtains a great **HighSchool_PR** score, he/she should keep up with the good work in order to perform well in **College_Score** three years later. On the other hand, if a student does not obtain a great **HighSchool_PR** score for any reason, he/she still has a second chance to do well in the **College_Score**.

[88]The high school is three years in Taiwan (grades 10-12).

# 10 Model Metrics: ROC and AUC

To measure the model performance in binary classification, we should evaluate the model's capability to distinguish between positive and negative datapoints.[89] For the model validation in Section 9.3, we have to choose a probability threshold in advance to convert the prediction results into the percentages. A predicted probability value is assigned a 1 (positive class) if it is above the threshold, and assigned to 0 (negative class) otherwise. Setting a low threshold allows us to get more predicted positive datapoints, but this increases the risk of getting false positives. On the contrary, setting a high threshold ensures that the predicted positive datapoints are more likely to be actually positive. But the downside is that many actual positive datapoints may not get classified as positives due to the high threshold.

Therefore, we introduce the ROC (receiver operating characteristic) as a model performance metric that summarizes over all possible probability thresholds.[90] The ROC plots the FPR (false positive rate) against the TPR (true positive rate), and the aggregate measure is called AUC (area under the curve).[91] AUC is between 0 and 1 like a probability. A model that is 100% correct has an AUC of 1, and a model that is 0% correct has an AUC of 0. However, if we know that a model is going to be 0% correct, we can infer that the actual datapoints are the complete opposite of the model's prediction. This gives us the same level of information as if the model had been 100% correct.[92] The baseline should be a completely random guess, where we get 50% correct like a coin flip. Hence the baseline AUC is 0.5.

According to Yang and Berdine (2017), the AUC values can be interpreted as below:

- AUC $< 0.5$: Performance worse than random guess.
- AUC $= 0.5$: Completely random guess, i.e., no classification power at all.
- $0.5 <$ AUC $< 0.6$: Unacceptable performance.
- $0.6 \leq$ AUC $< 0.7$: Minimally acceptable performance.
- $0.7 \leq$ AUC $< 0.8$: Adequate performance.
- $0.8 \leq$ AUC $< 0.9$: Great performance.
- AUC $\geq 0.9$: Excellent performance.

In non-clinical data science, AUC $\geq 0.7$ is typically required for a model to be effective (Han, 2022). For extremely difficult tasks, $0.6 \leq$ AUC $< 0.7$ may be tolerated (Kanter and Veeramachaneni, 2015). If the AUC is close to or less than 0.5, the model has little-to-no power in binary classification because the results are similar to a random guess.

The `R` package `verification` (Gilleland, 2015) is a tool to generate the ROC plot and calculate the AUC. The package was originally created to verify weather forecast data, and that's how it got the name `verification`. We will start with a demonstrative ROC curve in Section 10.1 using simulated data. Then in Section 10.2, we explore the data and model results, share our observations, and prepare the input for the ROC curve. Finally, we plot the ROC curve from our data and find the AUC in Section 10.3.

## 10.1 Demonstrative ROC Curve

We can create a demonstrative ROC plot with the `roc.plot` function in the `R` package `verification`. We are using data that consist of actual binary outcomes (0/1), along with the predictive probabilities of each datapoint to be 1. Then we will explore different features of the `roc.plot` function, and explain how to interpret the results.

[89] https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5
[90] https://towardsdatascience.com/understanding-the-roc-curve-and-auc-dd4f9a192ecb
[91] https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc
[92] https://victorzhou.com/blog/information-gain/

### 10.1.1 Simulated Data

Let's generate some random data as a hypothetical example, where the first 500 observations are 0 and the remaining 500 observations are 1. The vector array `obs_outcomes` is the actual observed binary outcomes. This array consists of elements only 0's and 1's.

```
library(verification)
group_size = 500
obs_outcomes = c(rep(0,group_size),rep(1,group_size))
```

We also assume a model: For an observed 0, the model predicts it as average 25% chance to be 1. For an observed 1, the model predicts it as average 75% chance to be 1. Let's generate the (fictitious) predicted probabilities using some statistical distributions.

- For an observed 0, we would like the predicted probabilities to be centered at 0.25, so we start with a normal distribution with mean 0.25 and standard deviation 0.3.
- For an observed 1, we would like the predicted probabilities to be centered at 0.75, so we start with a normal distribution with mean 0.75 and standard deviation 0.3.

However, a normal distribution may generate values outside the range of [0,1], so we add a uniform distribution to offset this issue. This adjustment ensures all predicted probabilities are valid, i.e., between 0 and 1.

- When the normal distribution outputs a value less than 0, we should replace the value with another value generated by the uniform distribution between 0 and 0.01.
- When the normal distribution outputs a value larger than 1, we should replace the value with another value generated by the uniform distribution between 0.99 and 1.

Now let's generate the predictive probabilities for the actual 0 and 1 datapoints, and we store them in the vector arrays `pred_prob_0` and `pred_prob_1`, respectively.

```
set.seed(3333)
pred_prob_0 = rnorm(group_size, mean=0.25,sd=0.3)
# Ensure all probability values are greater or equal to 0.
pred_prob_0 = pmax(pred_prob_0, runif(group_size,min=0,max=0.01))
# Ensure all probability values are less than or equal to 1.
pred_prob_0 = pmin(pred_prob_0, runif(group_size,min=0.99,max=1))

pred_prob_1 = rnorm(group_size, mean=0.75,sd=0.3)
# Ensure all probability values are greater or equal to 0.
pred_prob_1 = pmax(pred_prob_1, runif(group_size,min=0,max=0.01))
# Ensure all probability values are less than or equal to 1.
pred_prob_1 = pmin(pred_prob_1, runif(group_size,min=0.99,max=1))
```

In the code above, we use `set.seed` to specify a random seed to ensure reproducibility of outcomes.

- The functions starting with "r" (random) and a distribution name generates random values from the distribution given its parameters. For example, `rnorm` generates random values from a normal distribution, given its mean and standard deviation. Similarly, `runif` generates random values from a continuous uniform distribution, given its lower bound and upper bound. Note that `runif` is pronounced as "r-unif" instead of "run-if".
- In the functions `pmin` and `pmax`, the first letter "p" stands for "parallel" computation. Each function takes an input of two numerical vectors of the same length, and returns a single vector with the "parallel" minima (or maxima) of the input vectors. In other words, we compare the two input vectors to find the minimum (or maximum) for each position.

Under the scheme of combining the normal distribution and the uniform distribution, we can check that all output values are valid probabilities (i.e., between 0 and 1).

```
number_of_valid_prob_0 = sum((pred_prob_0 >= 0) & (pred_prob_0 <= 1))
number_of_valid_prob_1 = sum((pred_prob_1 >= 0) & (pred_prob_1 <= 1))

print(paste("Among all",group_size,"values in pred_prob_0,",
            number_of_valid_prob_0, "of them are valid probabilities."))
```

```
## [1] "Among all 500 values in pred_prob_0, 500 of them are valid probabilities."
```

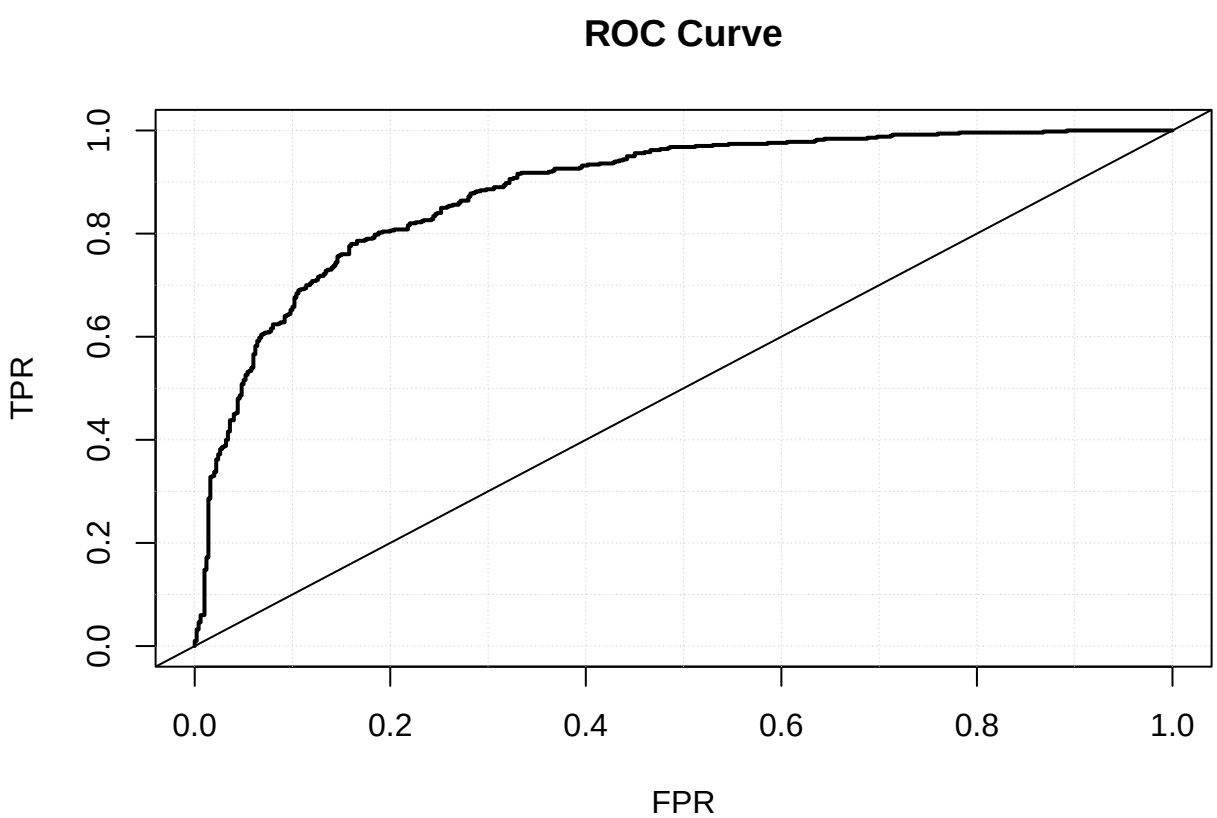
```
print(paste("Among all",group_size,"values in pred_prob_1,",
            number_of_valid_prob_1, "of them are valid probabilities."))
```

```
## [1] "Among all 500 values in pred_prob_1, 500 of them are valid probabilities."
```

### 10.1.2 The `roc.plot` Function

Now let's make the demonstrative ROC curve using `roc.plot`, which plots the false positive rate (FPR) against the false negative rate (FNR). The first input vector is the actual observed outcomes `obs_outcomes`, and the second input vector is the predicted probabilities `pred`. The curve should be above the straight line from the bottom left (0,0) to the top right (1,1). We also store the output of `roc.plot` in a new object `new_test` for future use, which is of `roc.data` type.

```
pred = c(pred_prob_0, pred_prob_1)
new_test = roc.plot(obs_outcomes,pred,xlab = "FPR", ylab="TPR",show.thres = FALSE)
```




We can access the model information in the `new_test` object with the `roc.vol` label.

```
print(new_test$roc.vol)
```

```
##      Model     Area      p.value binorm.area
```

```
## 1 Model  1 0.885536 3.501633e-99          NA
```

- The `Area` is the AUC (area under the curve), which is approximately 0.89. Typically we need AUC at least 0.7 for the model performance to be considered adequate. The straight line is the baseline as AUC = 0.5.
- The `p.value` is calculated against the null hypothesis $H_0$ as random guess (AUC $\approx$ 0.5). In this example, the p-value is extremely small, showing statistically significant evidence to reject $H_0$.
- The `binorm.area` is `NA` (not available). This applies only if we select a binormal fit.

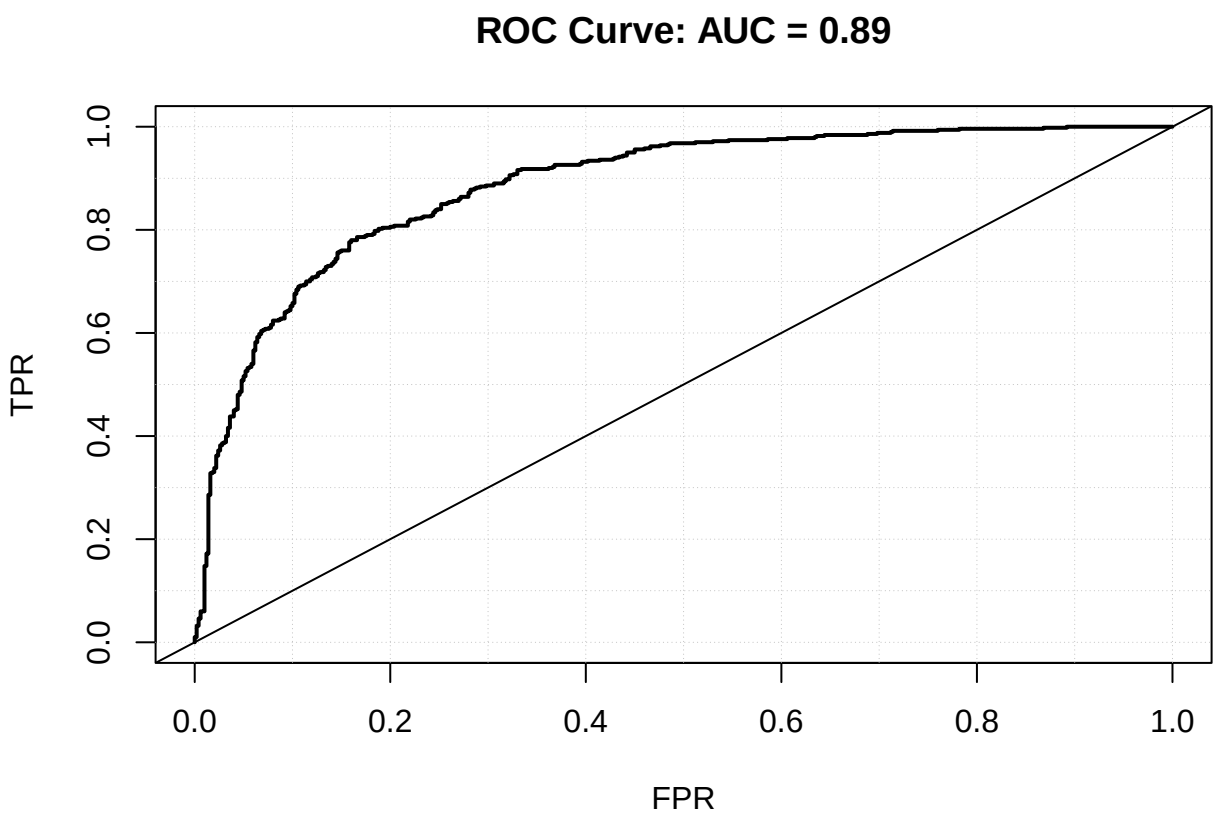
We can also directly retrieve the AUC value.

```r
auc_value = new_test$roc.vol$Area
print(auc_value)
```

```
## [1] 0.885536
```

Let's add the AUC value to the title of the ROC graph.

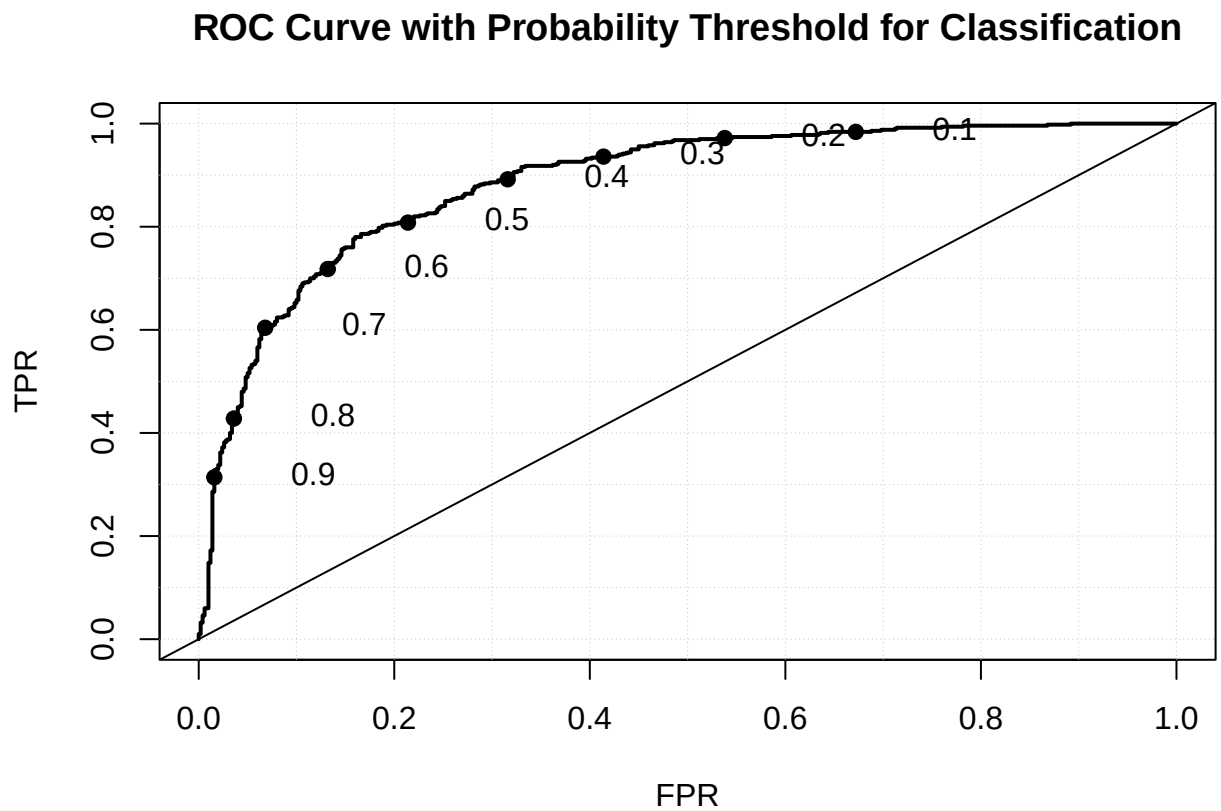
```r
roc.plot(obs_outcomes,pred,xlab = "FPR", ylab="TPR",show.thres = FALSE,
         main=paste("ROC Curve: AUC =", round(auc_value,2)))
```




### 10.1.3 More Details

The ROC graph is a summary of the model performance across all probability thresholds, and each point on the curve represents the FPR (false positive rate) and TPR (true positive rate) of a single probability threshold. When we set `show.thres = TRUE` in the `roc.plot` function, the graph shows that the probability thresholds range from 0 to 1.

```r
roc.plot(obs_outcomes,pred,xlab = "FPR", ylab="TPR",show.thres = TRUE,
         main="ROC Curve with Probability Threshold for Classification")
```

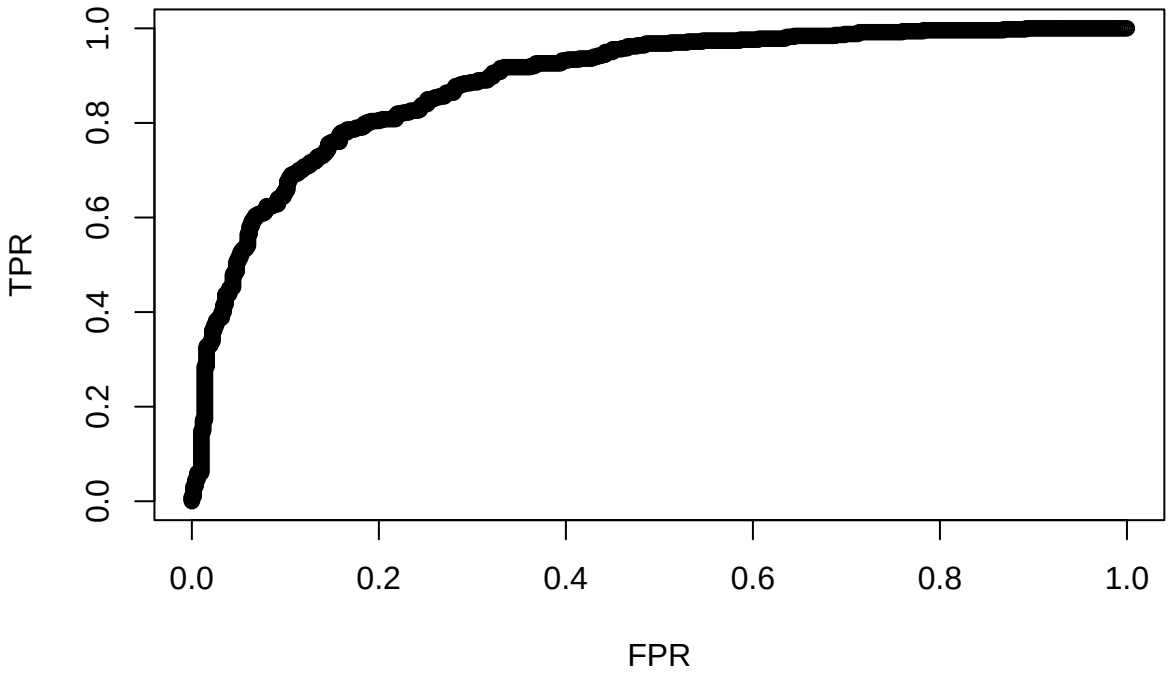



- At the left of the curve, a high probability threshold such as 0.9 sets a high criteria for the model to predict a datapoint as positive. This results in a low FPR , because when the datapoint is actually negative, the model is extremely unlikely to predict it as positive. But on the other hand, the high probability threshold also makes it relatively difficult to predict a datapoint to be positive. Hence the TPR is low because the actual positive datapoints may or may not receive a positive prediction.
- At the right of the curve, a low probability threshold such as 0.1 sets a low criteria for the model to predict a datapoint as positive. This results in a high TPR, because when the datapoint is actually positive, the model is extremely likely to predict it as positive. But on the other hand, the low probability threshold also predicts too many datapoints to be positive, and some of them are actually negative. Hence the FPR is also high.

The underlying data for the ROC graph is stored in the matrix `new_test$plot.data`. The first column (`V1`) is the probability threshold used to convert predictive probabilities into binary predictions. The second column (`V2`) is the empirical hit rate (true positive rate, TPR), and the third column (`V3`) is the false alarm rate (false positive rate, FPR).

We can manually plot the FPR against the TPR, but this is NOT recommended. We should use a pre-build function whenever possible, because these functions have been optimized for performance. The ROC graph we created is not as clean as the one generated by the function `roc.plot` in Section 10.1.2.

```r
test_df = as.data.frame(new_test$plot.data)
plot(test_df$V3, test_df$V2, xlab="FPR", ylab="TPR")
```




## 10.2 Data Observation and Processing

At the beginning of Section 8.1, we used 0.5 as the default probability threshold to classify whether a student would obtain **College_Score** at least 65 or not. That is, the predicted probability needs to be 0.5 or higher for a datapoint to be predicted as positive (i.e., getting **College_Score** at least 65). This is acceptable because the data are balanced and 48.9% of students in the data made the cut, i.e., obtained **College_Score** at least 65. We are doing this project as an experiment, without trying to optimize any particular metric. But the probability threshold of 0.5 may not be appropriate in imbalanced datasets. If a dataset consists of 80% samples in the majority category and 20% samples in the minority category, the threshold to predict the minority category would be closer to 80% to maximize the accuracy.[93] Plus, depending on our goal to maximize recall/prediction/other metrics, we may choose a different threshold for the probability-to-decision conversion.

In many situations, it is not obvious to determine the probability threshold to predict a datapoint to be positive. A higher probability threshold would result in fewer points to be predicted as positive, so we may not find all of the true positive datapoints. But this model may be better at classifying true negative datapoints as negative, since the requirements are high to predict a datapoint to be positive. On the other hand, a lower probability threshold would predict more points as positive, increasing the chances of capturing all of the true positive datapoints. But this may also result in many true negative datapoints being predicted as positive. Hence there is a tradeoff between recall and precision when we select the probability threshold.

Let's revisit the leave-one-out cross-validation results from Section 9.2.2, and explore how the probability threshold changes the accuracy/precision/recall/FPR/FNR. Before deciding on the threshold, we need to look at the boundary conditions – the minimum and maximum predicted probabilities in the data. If the threshold is below the minimum, the model will predict all points to be true, which is not useful. If the

[93] https://towardsdatascience.com/tackling-imbalanced-data-with-predicted-probabilities-3293602f0f2

threshold is above the maximum, the model will predict all points to be false, which is not useful, either. In the array of predictive probabilities `prob_leave1out`, the minimum is 0.0025 and the maximum is 0.7782. We should choose a probability threshold between these two values.

```
check_prob_leave1out = as.numeric(prob_leave1out)

print(paste("Min prob_leave1out:", min(check_prob_leave1out)))
```

```
## [1] "Min prob_leave1out: 0.00249702111740674"
```

```
print(paste("Max prob_leave1out:", max(check_prob_leave1out)))
```

```
## [1] "Max prob_leave1out: 0.778173314094346"
```

When the probability threshold is set to 0.5 as the default, the precision is 67% and the recall is 78%. We use these numbers as a comparison baseline. We also show the confusion matrix, so that the readers can observe the patterns from different probability thresholds.

```
original_leave1out =  prob_to_matrix(data_corr, prob_leave1out, threshold=0.5)
print(original_leave1out)
```

```
##                 test_pred_65up
## test_actual_65up TRUE FALSE
##            TRUE    72    20
##            FALSE   35    61
```

```
original_results = confusion_to_measures(original_leave1out)
print(round(original_results, digits=4))
```

```
##  Accuracy Precision    Recall       FPR       FNR
##    0.7074    0.6729    0.7826    0.3646    0.2174
```

When the probability threshold is 0.7, the precision increased to 82%, but the recall decreased to 52%. A higher probability threshold means the model is less likely to predict a datapoint to be True, so a positive prediction is more likely to be actually positive, increasing the precision. However, the model may miss more datapoints (i.e., predict as negative) which are actually positive, resulting in a drop in the recall.

```
high_leave1out = prob_to_matrix(data_corr, prob_leave1out, threshold=0.7)
print(high_leave1out)
```

```
##                 test_pred_65up
## test_actual_65up TRUE FALSE
##            TRUE    48    44
##            FALSE   10    86
```

```
high_results = confusion_to_measures(high_leave1out)
print(round(high_results, digits=4))
```

```
##  Accuracy Precision    Recall       FPR       FNR
##    0.7128    0.8276    0.5217    0.1042    0.4783
```

When the probability threshold is 0.3, the recall increased from 78% to 89%, but the precision decreased from 67% to 59%. A lower probability threshold means the model is more likely to predict a datapoint to be True, so the model has a better chance of catching most of the datapoints which are actually positive, increasing the recall. But the drawback is that when the model predicts a datapoint to be True, the datapoint has a greater risk of not actually being positive. In other words, using a low threshold may result in more false positives, hence reducing the precision.

```
low_leave1out = prob_to_matrix(data_corr, prob_leave1out, threshold=0.3)
print(low_leave1out)
```

```
##                 test_pred_65up
## test_actual_65up TRUE FALSE
##            TRUE    82    10
##            FALSE   56    40
```

```
low_results = confusion_to_measures(low_leave1out)
print(round(low_results, digits=4))
```

```
##  Accuracy Precision    Recall       FPR       FNR
##    0.6489    0.5942    0.8913    0.5833    0.1087
```

## 10.3 Implementation and Results

We decided to use the leave-one-out results to plot the ROC-AUC, rather than the results from K-fold cross-validation or separate training and testing datasets. In the outcomes of separate training and testing datasets (Section 9.1), only the testing dataset contains predictive probabilities, and we do not think the sample size is large enough for the overall evaluation of model performance. In comparison, both leave-one-out and K-fold cross-validation predict every single record in the data. The main difference is that in leave-one-out (Section 9.2.2), each datapoint is predicted by a different training sample. When two datapoints have the same **HighSchool_PR** but differ in **College_Score**, they will have different predictive probabilities because the associated training samples are not the same. On the other hand, K-fold (Section 9.2.1) has only $K$ different subsets, so two datapoints in the same subset with the same **HighSchool_PR** will get the same predicted probability for **College_Score**. This can result in many repetitive predicted values, which is impractical in real life.

### 10.3.1 Original Data

In the implementation of ROC-AUC curve, we use the leave-one-out cross validation output because it predicts every single record in the data. In the first `roc.plot` input, we use `plot=NULL` to avoid creating a plot prematurely. From the `roc.plot` object `roc_data_test`, we can retrieve the AUC value, which is approximately 0.79. This AUC means our model has good performance, according to the guidelines at the beginning of Chapter 10.

```
library(verification)
set.seed(1234)

real_outcomes = data_corr$College_Score >=65
pred_outcomes = as.numeric(prob_leave1out)

roc_data_test = roc.plot(real_outcomes, pred_outcomes,
                         xlab = "FPR", ylab="TPR", show.thres = FALSE,
                         plot=NULL)

auc_value_test = roc_data_test$roc.vol$Area
print(auc_value_test)
```

```
## [1] 0.7898551
```

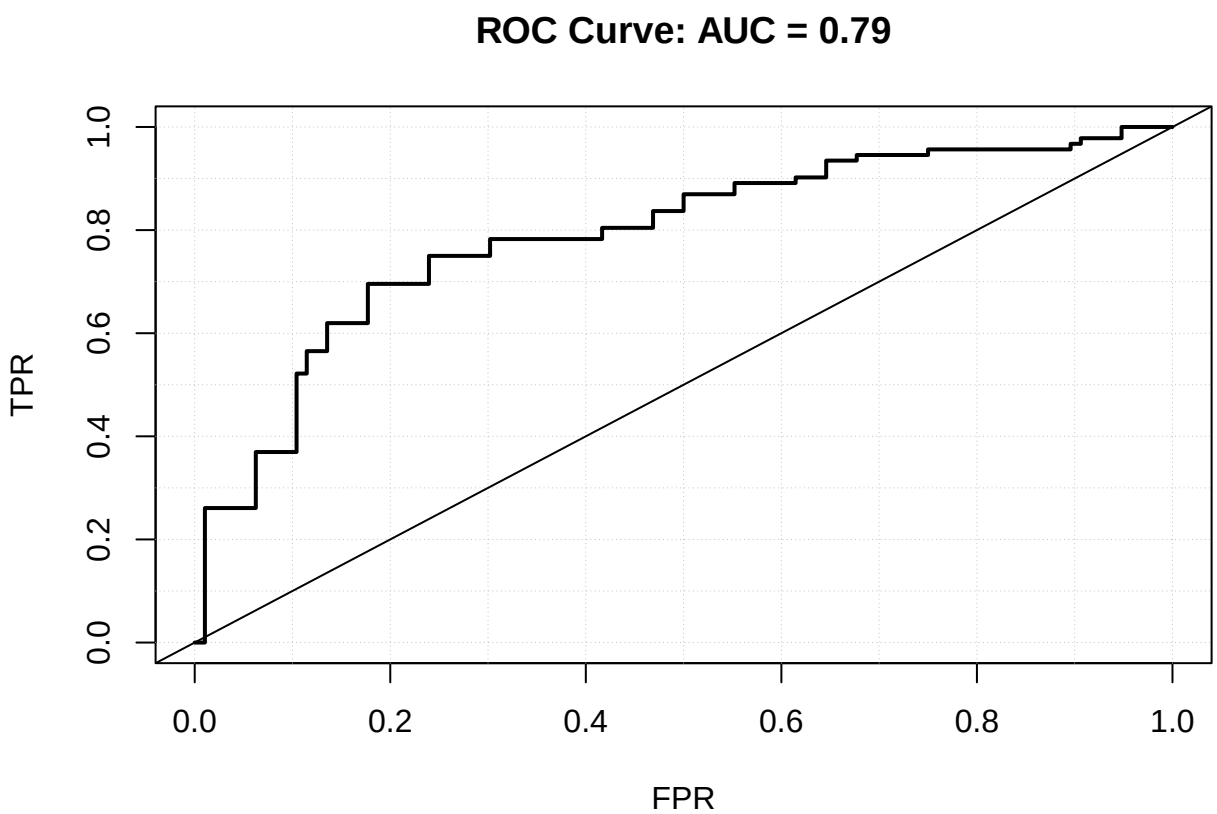
Now we add the AUC value to the title, and finally plot the ROC graph from our data. However, this ROC curve contains "steps" and is not smooth, because our model made predictions from discrete data with a relatively small sample size.[94] Each student with the same **HighSchool_PR** should have the same predictive probability of getting **College_Score** at least 65, and the repetitive predictive probablities resulted in the discontinuous steps in the graph.

Hence we need to smooth the ROC curve. The `R` package `pROC` (Robin et al., 2011) contains many methods to smooth the ROC curve. But `pROC` generates the ROC curve based on specificity and sensitivity, rather than

[94] https://forum.posit.co/t/why-do-smooth-roc-curves/121753

FPR and TPR. If we still use the R package `verification`, an easier solution is to manually add a small random noise to **HighSchool_PR** in the data, i.e., jittering. In this way, each subject's **HighSchool_PR** becomes slightly different than each other, so that we can get distinct values.

```
roc.plot(real_outcomes, pred_outcomes,
         xlab = "FPR", ylab="TPR",  show.thres = FALSE,
         main=paste("ROC Curve: AUC =", round(auc_value_test,2)))
```




### 10.3.2 Jittered Data

Jittering is the process of adding some noise to the numbers, and we can use the R function `jitter` perform this modification. The input parameter `factor` controls the variance of the noise. For example, we jitter the numbers 1, 1, 2, 2, 3, 3. Using `factor=2` gives a larger amount of noise than using `factor=1`. Observe that jittering the same value twice generates separate versions of random noise, so the two output numbers are slightly different. We should be careful not to add too much noise, otherwise the original data signal will be buried by the noise. We also need to specify the random seed to ensure reproducibility.

```
example_vector = c(1,1,2,2,3,3)
set.seed(20)
jitter(example_vector, factor=1)
```

```
## [1] 1.151009 1.107413 1.911585 2.011665 3.185163 3.192142
```

```
example_vector = c(1,1,2,2,3,3)
set.seed(20)
jitter(example_vector, factor=2)
```

```
## [1] 1.302017 1.214827 1.823171 2.023331 3.370326 3.384284
```

Before we start jittering the data, we also need to augment the data because we only have 188 valid records in the sample. Data augmentation is the process of artificially generating new data from the original data, mainly for machine learning training purposes.[95] Therefore, we duplicate the 188-record dataset by 20 times, resulting in 3760 total rows. Then we jitter all 3760 datapoints of **HighSchool_PR**, using `factor=2` for the level of noise. An earlier example showed that jittering the same number can produce different outputs, so we can rest assured that each jittered value of **HighSchool_PR** is distinct. In the code, we can see that the first two datapoints are approximately 60.3 and 60.2, indicating that both jittered values had an original **HighSchool_PR** of 60. Parameters are empirically determined here.

```
# Augment the dataset to 188*20=3760 rows.
augment_factor = 20
data_jitter = data_corr
for (aa in 2:augment_factor) {
  data_jitter = rbind(data_jitter, data_corr)
}

# Use a larger variance for the jitter.
set.seed(20)
data_jitter$HighSchool_jitter = jitter(data_jitter$HighSchool_PR, factor=2)

# Show the first 10 jittered values
print(data_jitter$HighSchool_jitter[1:10])
```

```
##  [1] 60.30202 60.21483 98.82317 92.02333 87.37033 69.38428 97.67307 94.65660
##  [9] 87.86208 98.89606
```

**Remark**: `R` uses pass by value in data assignment, so the whole content in `data_corr` is copied to `data_jitter`. Changing the data in `data_jitter` won't change the original values in `data_corr`, which is good. Compared with the pandas DataFrame in Python, we have to explicitly use the function `copy` to make a deep copy of the data. If we simply use assignment, the new DataFrame is simply a pointer to the original DataFrame. Then editing the new dataset will result in changing the original data.[96]

Now let's perform the leave-one-out cross validation on the jittered version of data, which contains 3760 total rows. The approach is exactly the same as described in Section 9.2.2. We reserve one record for testing, and use all the other records to train the model. Then the model predicts the one datapoint which was left out before. Given the size of the data, it may take more than a few seconds to complete this part and obtain the predictive probabilities.

```
# Leave-one-out for the jittered version of data

nn = nrow(data_jitter)
# 188*20=3760 total rows of data

prob_leave1out_jitter = rep(c(-1), nn)

for (ii in 1:nn) {
  data_test = data_jitter[ii,] # reserve one record for testing
  data_exclude = data_jitter[-ii,]

  train_leave1out = glm(CS_65up ~ HighSchool_jitter, data=data_exclude, family="binomial")
  # summary(train_leave1out)

  test_leave1out = predict.glm(train_leave1out, data_test, type="response")
  # type="response" gives the predicted probabilities
```

[95] https://aws.amazon.com/what-is/data-augmentation/
[96] https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.copy.html

```
  # Store the predicted probability to the general list
  prob_leave1out_jitter[ii] = test_leave1out
}
```

In `roc.plot`, we need to specify the probability thresholds. Otherwise, the default in `roc.plot` is to use all unique thresholds for each prediction, and there will be way too many unique probability thresholds given a large dataset. We will get the warning message: "Warning: Large amount of unique predictions used as thresholds. Consider specifying thresholds."

Hence we specify the thresholds as 0.001, 0.002, . . . , all the way to 0.999 and 1.000. There are 1000 thresholds in increments of 0.001 within the interval (0,1], so we can create this numerical vector using `c(1:1000)/1000`. The number 1000 is called the `granularity`.[97]

Although the AUC is slightly higher using the jittered data (81%) than the original data, the small difference is due to the random noise from jittering. If we select another random seed, the resulting AUC would also be close to 80%.

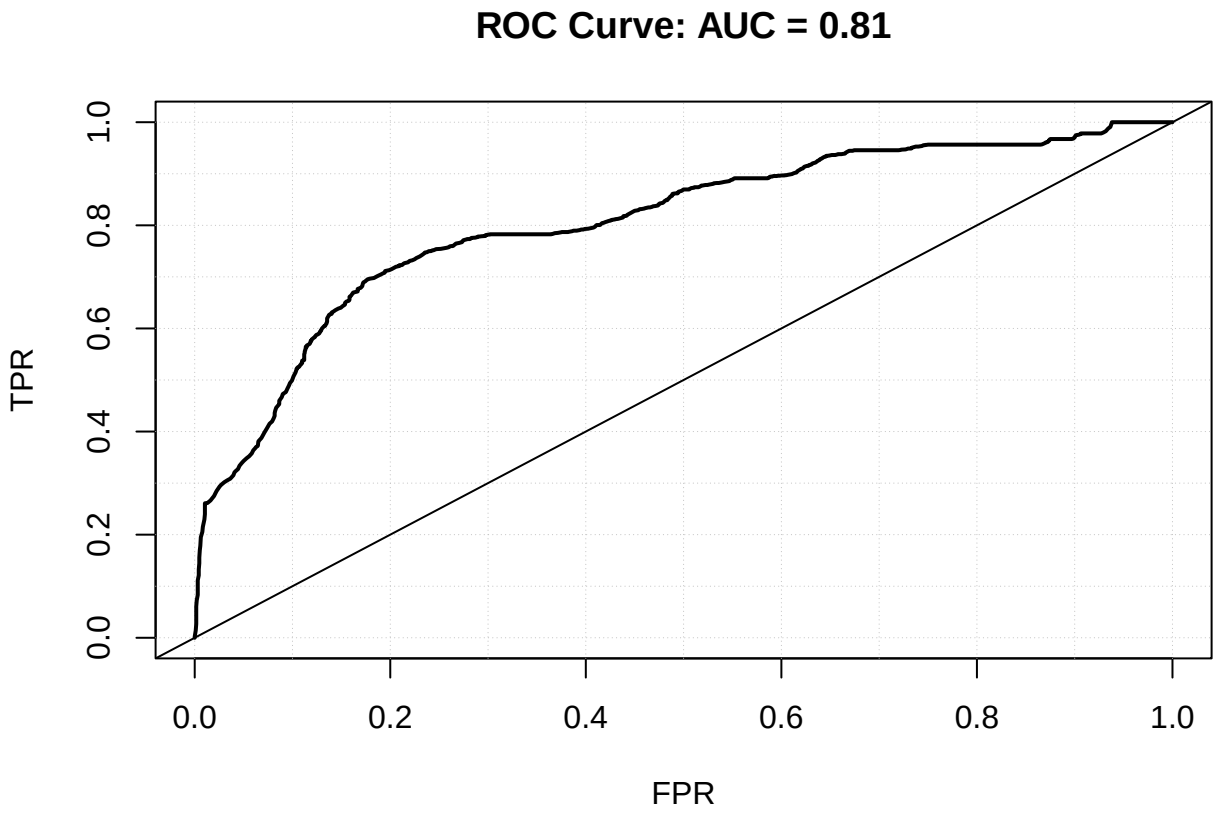
```
library(verification)
set.seed(1234)

real_outcomes_jitter = data_jitter$College_Score >=65
pred_outcomes_jitter = as.numeric(prob_leave1out_jitter)

# Probability thresholds: 0.001, 0.002, ..., 0.999, 1.000.
granularity = 1000
thresholds = c(1:granularity)/granularity

roc_data_jitter = roc.plot(real_outcomes_jitter, pred_outcomes_jitter,
                           thresholds = thresholds,
                           xlab = "FPR", ylab="TPR", show.thres = FALSE,
                           plot=NULL)

auc_value_jitter = roc_data_jitter$roc.vol$Area

print(roc_data_jitter$roc.vol$Area)
```

```
## [1] 0.8056912
```

Now the ROC graph is smoother than the original one in Section 10.3.1.

```
roc.plot(real_outcomes_jitter, pred_outcomes_jitter,
         thresholds = thresholds,
         xlab = "FPR", ylab="TPR",  show.thres = FALSE,
         main=paste("ROC Curve: AUC =", round(auc_value_jitter,2)))
```

[97] https://www.talon.one/glossary/granularity




## 11 Recap of the Project

This manuscript is a detailed example of a project from data collection to meaningful insights, and we assume a background equivalent to Statistics 101. In the end-to-end execution, we also demonstrated good coding practices such as user-defined functions, descriptive variable names, and helpful comments in code. These practices in `R` are transferable to other programming languages like Python and Java, which are commonly-used in the tech industry.

We obtained small-scale data that are publicly available and articulated our thought process in each step. We started with data preprocessing (i.e., cleaning the data), explored the data, and applied statistical models to quantify the relationship between the two variables of interest. The first model we try does not always work out, and the model validation can reveal more data issues previously undiscovered in the exploratory phase. Given the new findings, we can try different ways to continue the analysis.

Our original problem is to find the relationship between high school entrance exam scores and college entrance exam scores. The exploratory analysis in Chapter 4 shows a moderate positive association between the two sets of scores, with the correlation coefficient 0.50. However, this does not necessarily mean that the linear regression is the most appropriate model. Linear regression is widely used, but it is not a panacea for data analysis. Chapter 5 states the four assumptions to be met in linear regression: linearity, nearly normal residuals, constant variability, and independent observations. When these assumptions are violated, we should not use the linear regression model.

Although the linear regression did not work out, we took a deep dive to understand the data breakdown by high school entrance exam score category in Chapter 6. It is possible that a model has different levels of performance in different ranges of the independent variable. Exploratory data analysis shows the data are left-skewed, i.e., much more respondents with high scores than low scores. This chapter is focused on the top scorers because they account for the majority of the respondents.

The good news is that we can continue the data analysis through several ways:

- Transform the data to meet the assumption requirements
- Choose another model to investigate the relationship between the two variables
- Reformulate the problem and answer a different question from the data

In Chapter 7, we reformulated the problem to: Given a student's high school entrance exam score, estimate the probability of getting a college entrance exam score of at least 65. This statement binarizes the outcome of college entrance exam scores, so we discovered the relationship of the two variables in a different way. Sections 7.3 and 7.4 show that when the high school entrance exam score increases by one percentile, the odds of "success" increases by about 16.1% on average, i.e., getting a college entrance exam score at least 65. We are 95% confident that the odds increase is between 10.6% and 22.9%.

After implementing the logistic regression model, we also validated the results using in-sample prediction (Chapter 8) and out-of-sample prediction (Chapter 9). The accuracy is approximately 70% and consistent across both methods of prediction. About half of the respondents obtained a college entrance exam score at least 65, so the baseline is around 50% like a coin-flip. Hence the logistic regression model is doing much better compared with the baseline, and the detailed metrics are listed and explained in Section 9.3.

We also examined the model results by high school entrance exam score category in Section 8.3. Since this is a binary classification model, we focused on the confusion matrices of each category. For the respondents whose high school entrance exam score in the 0-79th percentile, the model predicted none of them to obtain a college entrance exam score of at least 65, but some respondents actually did succeed. For the respondents in the 80-89th percentile, the predictions are the same but with a higher proportion of unexpected successes. The model predictions consist of both success and non-success outcomes for the respondents in the 90-94th percentile. Finally, the model predicts all respondents in the 95-99th percentile to obtain a college entrance exam score of at least 65, and inevitably some of them did not achieve this. Given the single outcome prediction in most categories, we added zero columns/rows to ensure the 2x2 size of each confusion matrix.

# 12 Recommended Resources for Learning

We would like to direct readers to resources for learning statistics and data science, in order to expand their professional skillsets. Online information is abundant and probably overwhelming, so we decided to categorize some examples based on difficulty level and the prerequisites. We also provide a brief description of each resource including goals and/or applications, so that readers can make informed decisions in choosing their own professional development activities. Section 12.1 points to some materials for learning statistics, while Section 12.2 takes a step forward to data science and programming.

## 12.1 Resources for Learning Statistics

Statistics is the core of data analysis.[98] Even with the aid of machine learning algorithms and integrated computation tools, people still need to know the fundamental concepts in statistics to appropriately interpret the numbers from the data. For instance, when you take a sample from the population data, how do you verify that the sample is a representative sample? How do you interpret the model coefficients in the output? How do you ensure that all predictive probabilities are between zero and one? Integrated tools make it easy to build and run a model, but what's more important is to choose an appropriate model to implement for the project goal. We also need to be able to detect and troubleshoot issues in the model output.

Here are some resources we recommend for the path of learning statistics, from introductory to advanced. Although some readers may be more interested in data science, learning statistics serves as a building block to becoming an expert data scientist. Human decisions also play a key role in getting useful insights from data,[99] and the statistical skills would help readers in making better decisions in data.

[98] https://datascience.virginia.edu/news/how-much-do-data-scientists-need-know-about-statistics
[99] https://ww2.amstat.org/meetings/sdss/2020/onlineprogram/AbstractDetails.cfm?AbstractID=308230

The Statistics 101 course covers the fundamental statistical concepts for students to perform data analysis, and the programming component is often included. For readers who would like a refresher, we recommend the textbook *OpenIntro Statistics* (Diez et al., 2019) and the online course series *Statistics with R Specialization* on Coursera.[100] Basic concepts like hypothesis testing and confidence intervals are also asked in technical interviews for data science jobs. Beware! For an overview of Bayesian Statistics, we recommend the *Bayesian Statistics* online course on Coursera,[101] along with the open-source textbook *An Introduction to Bayesian Thinking* (Clyde et al., 2018). At this stage, knowledge of calculus is helpful but not required.

As the next step beyond the introductory level, we suggest reading *The Statistical Sleuth: A Course in Methods of Data Analysis* (Ramsey and Schafer, 2013), which is the textbook for undergraduate-level regression analysis at Duke Statistical Science.[102] The book covers intermediate topics such as ANOVA (Analysis of Variance) and multiple linear regression. It also provides data files for case studies and exercises.[103] *Statistics: Unlocking the Power of Data* (Lock et al., 2020) is heavily focused on data analysis with real applications, and this textbook introduces computer simulation methods like bootstrap intervals and randomization tests for statistical inference. For a more theoretical overview of mathematical statistics, we recommend *Probability and Statistics* (DeGroot and Schervish, 2012) where calculus is a prerequisite. A solid theoretical background helps readers bridge the gap between writing code and understanding complex statistical models. When the readers go further in statistics, college-level mathematics becomes increasingly important. Formulas and equations will gradually appear in the curriculum, especially calculus for continuous functions and matrix algebra for discrete datapoints.

For the advanced readers, we recommend the following graduate level statistics textbooks. Note that multivariate Calculus and matrix algebra are absolutely necessary at this stage. *A First Course in Bayesian Statistical Methods* (Hoff, 2009) explains in detail on how to perform Bayesian statistical modeling, which is much more comprehensive than the Bayes' rule (conditional probability). For a focus on linear models, we suggest *A Modern Approach to Regression with R* (Sheather, 2009) with solid examples, especially on model validation and residual analysis. *Statistical Inference* (Casella and Berger, 2021) is a must-read if you are interested in the theory behind statistical concepts such as maximum likelihood estimation. These textbooks build on statistical concepts that readers may have already learned at the beginner to intermediate levels. If we can recommend only one textbook, the one will be *Categorical Data Analysis* (Agresti, 2003). This book explains many generalized linear models in detail, such as logistic regression, Poisson log-linear model, and proportional hazards for survival analysis models. There are lots of categorical data in real life, making these models particularly useful.

There are many other high-quality statistics textbooks in the universe, and we selected these as a starting point. If the readers are interested in a data science career, learning statistics builds the foundations and expands your toolbox to generate insights from data.[104] We also recommend exploring Kaggle datasets[105] for hands-on practice. To get closer to analyzing real-world data, we encourage the readers to attend data hackathons, contribute to open-source projects, and/or volunteer technical skills at nonprofit organizations like Statistics Without Borders.[106] These activities are not a comprehensive list, and please feel free to get creative and carve your own professional path forward.

## 12.2 Resources for Data Science and Programming

Since we discussed plenty of statistics resources in the previous section, we would like to provide a few pointers to **data science**, one of the hottest fields in tech. Programming is also essential to handle data, so we introduce some practical tools that readers can incorporate them into their data science workflow. (Please don't get overwhelmed with the massive selection of tools; you can try one at a time.)

[100] https://www.coursera.org/specializations/statistics
[101] https://www.coursera.org/learn/bayesian
[102] https://www2.stat.duke.edu/courses/Fall18/sta210.001/
[103] http://www.statisticalsleuth.com/
[104] https://towardsdatascience.com/the-difference-between-data-science-and-statistics-168c7062c201
[105] https://www.kaggle.com/datasets
[106] https://www.statisticswithoutborders.org/

This document is created with R Markdown (Allaire et al., 2019),[107] which has a framework to directly incorporate code and graphs in the report. This not only reduces the errors from copy-pasting tables and plots, but also ensures reproducibility of the data analysis (Baumer et al., 2014). R Markdown can generate reports in PDF, HTML, or Microsoft Word formats. R Markdown also provides a template for presentation slides, and the output can be in PDF, HTML, or Microsoft PowerPoint files. Note that PDF output from R Markdown requires the installation of LaTeX.[108] (An alternative way is to save the HTML output to a PDF, but this is a manual step outside the end-to-end pipeline.) For a full book with multiple chapters, we recommend using the R package `bookdown` (Xie, 2016) for a comprehensive and reproducible workflow.

The graphs in this manuscript are mainly for demonstration; they are created using the basic `plot` function as the default in RStudio. For advanced data visualization, we highly encourage readers to use the R package `ggplot2` (Wickham, 2016). This package provides much more capability to convey the message in a graph, and the foundations behind `ggplot2` are based on *The Grammar of Graphics* (Wilkinson, 2013). For a simple and hands-on guide for data visualization, we recommend the book *Visualize This* (Yau, 2011) to determine which types of graphs are most appropriate for which scenarios. On the other hand, *Semiology of Graphics: Diagrams, Networks, Maps* (Bertin, 1983) is a comprehensive textbook for data visualization theory. Concepts such as human perception are technology-agnostic; the principles apply to almost any tool for data visualization.

The R code is also for demonstration. The author makes every effort to ensure correctness and readability of the code, but errors and inconsistencies are inevitable. Since this is an open-source project on GitHub,[109] readers are welcome to make a pull request and/or report any issues. To learn more about better coding styles, we recommend the computer science books *Clean Code* (Martin, 2009) and *Code Complete* (McConnell, 2004). For complex operations in R, we recommend the R package `tidyverse` (Wickham et al., 2019) because it provides a fast, efficient, and organized workflow for general data modeling.[110] The book R *for Data Science: Import, Tidy, Transform, Visualize, and Model Data* (Wickham and Grolemund, 2016) demonstrates an end-to-end cycle of using `tidyverse` in data science.

In addition to R, Python is another commonly-used programming language in data science. Python is an object-oriented language which is closer to software development, while R is more focused on statistical models.[111] Our suggestion is to learn both R and Python, so that the readers get to choose the one that fits their unique needs. (The author does not have a strong personal preference, so she often chooses the one that her collaborators use to better leverage their existing code.) After you know a programming language, it would not be difficult to learn a new one because many concepts are similar, such as `for` loops and user-defined functions. Don't worry!

# 13 Personal Scores and Remarks

This manuscript is created because the author reflected on her own experience regarding high school and college entrance exams. Instead of dwelling on the past, the author decided to redirect her energy to something more meaningful and constructive. She leveraged the opportunity to apply her data science and research skills. In addition to statistical analysis, she also wanted to demonstrate full reproducibility of the project. Inspired by Baumer et al. (2014), the author decided to incorporate R Markdown as a reproducible analysis tool into introductory statistics. She also designed her own template to create LaTeX Beamer PDF slides from R Markdown,[112] in which she included some commonly-used LaTeX functions.

The author scored **HighSchool_PR** 95 in Year 2004, but she got admitted to Taipei First Girls' High School because the special class for math and science track had a separate admission process at that time. Taipei First Girls' High School[113] usually requires **HighSchool_PR** 99 for admission, with few exceptions

[107] https://rmarkdown.rstudio.com/articles_intro.html
[108] https://bookdown.org/yihui/rmarkdown/pdf-document.html
[109] https://github.com/star1327p/Exam-Scores-PTT
[110] https://www.r-bloggers.com/2018/09/why-learn-the-tidyverse/
[111] https://www.datacamp.com/tutorial/r-or-python-for-data-analysis
[112] https://github.com/star1327p/r-beamer-template
[113] http://web.fg.tp.edu.tw/~tfghweb/EnglishPage/index.php

such as recruited athletes[114] and students with disabilities.[115] Several years later, the policy changed so that students had to be admitted to the high school first in order to try for the special class placement within. Most students in the special class had **HighSchool_PR** 99 anyway, and the author was one of the few who did not. Hence the author wondered whether she would have similar academic performance, had she enrolled in a different high school commensurate to **HighSchool_PR** 95 in Taipei, Taiwan.

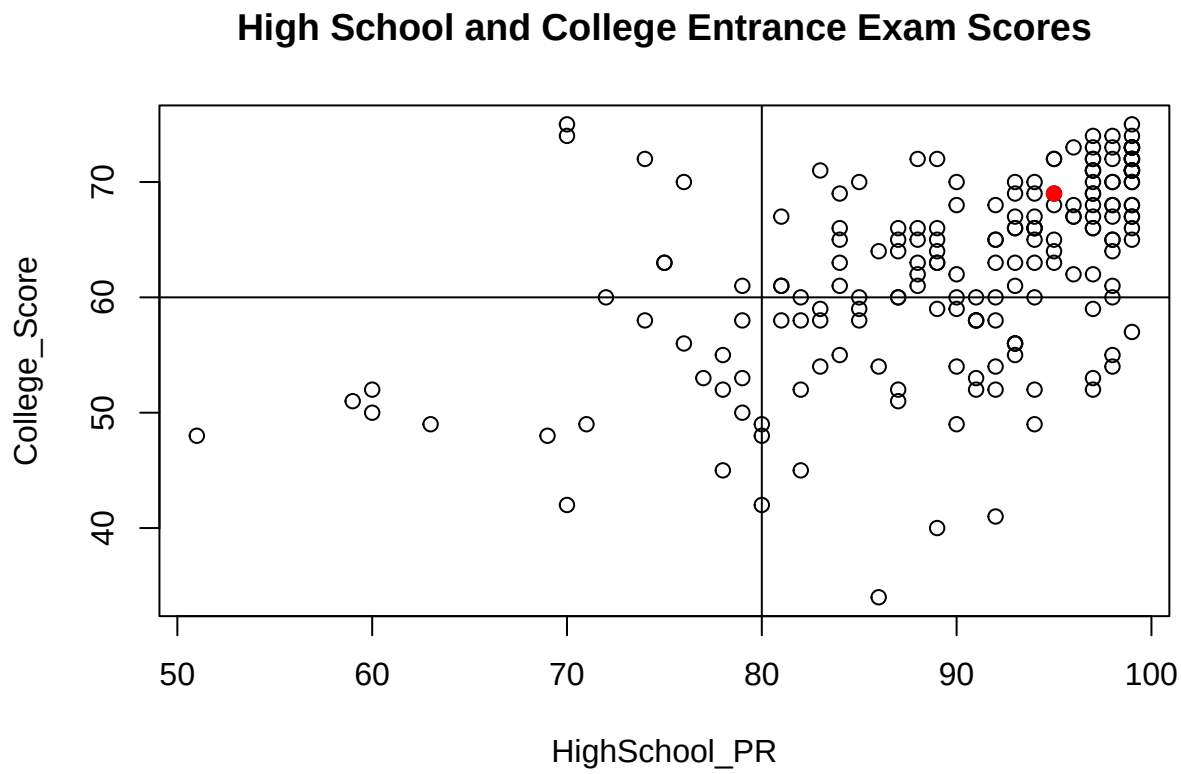
The author got **College_Score** 69 on the General Scholastic Ability Test (GSAT) in Year 2007. She applied to several colleges, and one of them offered her early admission.[116] However, she was not satisfied with the outcome, so she decided to try for the Advanced Subjects Test (AST). She obtained an excellent score on the AST and got admitted to The Department of Electrical Engineering at National Taiwan University (NTUEE), one of the top colleges in Taiwan. NTUEE[117] typically requires full marks (15 out of 15) in English, mathematics, and science in **College_Score** for the early admission through GSAT.[118] Most students at NTUEE had a **College_Score** of 70 or higher, at the time when 75 was the max possible score. But still a significant number of students got admitted through the AST in July, regardless of their GSAT score.

## 13.1 Comparison with the Data

The author's high school and college entrance exam scores are indicated as the red dot in the scatterplot below from Section 4.3. We also added the reference lines for **HighSchool_PR** 80 and **College_Score 60**. **HighSchool_PR** 95 and **College_Score** 69 are definitely above-average scores, despite imperfect.

```
plot(data_corr$HighSchool_PR, data_corr$College_Score,
     main = "High School and College Entrance Exam Scores",
     xlab="HighSchool_PR",
     ylab="College_Score")

abline(h=60,v=80)
points(x=95, y=69, col="red", pch=19)
```

[114] https://www.sa.gov.tw/PageContent?n=1265
[115] http://www.rootlaw.com.tw/LawArticle.aspx?LawID=A040080080001900-1020822
[116] https://sites.google.com/site/christinepeijinnchai/resume-cv
[117] https://web.ee.ntu.edu.tw/eng/index.php
[118] https://university.1111.com.tw/univ_depinfo9.aspx?sno=100102&mno=520101




Let's examine all of the **College_Score** datapoints given **HighSchool_PR** 95. There are seven points in the data, with median 68 and mean about 67.6. One of them is **College_Score** 69, the same score as the author's. (She personally did not respond to this survey in 2015 because she discovered this announcement a few years later.)

```
sort(data_corr$College_Score[which(data_corr$HighSchool_PR == 95)])
```

```
## [1] 63 64 65 68 69 72 72
```

Let's also inspect the nearby values, i.e., the **College_Score** values given **HighSchool_PR** 94 or 96.

For **HighSchool_PR** 94, there are ten **College_Score** values in the data. The median is 66, and the mean is 63. The author's **College_Score** would be the second highest score in this batch.

```
sort(data_corr$College_Score[which(data_corr$HighSchool_PR == 94)])
```

```
##  [1] 49 52 60 63 65 66 66 66 67 69 70
```

For **HighSchool_PR** 96, there are five **College_Score** values in the data. The median is 67, and the mean is 67.4. The author's **College_Score** would still be the second highest in this batch.

```
sort(data_corr$College_Score[which(data_corr$HighSchool_PR == 96)])
```

```
## [1] 62 67 67 68 73
```

Finally, let's explore the **College_Score** values of **HighSchool_PR** 97-99. There are 58 datapoints, which account for 30.8% of the total data. The median of **College_Score** is 70, and the mean is 68. Note that the max possible **College_Score** is 75, and few respondents achieved 74 or 75 in the entire dataset. In the earlier graph of this subsection, **College_Score** 69 is about in the middle of the **College_Score** datapoints for **HighSchool_PR** 97-99. This observation is similar to what we discovered in the subset.

```r
pr97to99 = sort(data_corr$College_Score[which(data_corr$HighSchool_PR %in% c(97,98,99))])

print(pr97to99)
```

```
##  [1] 52 53 54 55 57 59 60 61 62 64 65 65 65 66 66 66 67 67 67 67 68 68 68 68 68
## [26] 69 69 70 70 70 70 70 71 71 71 71 71 71 71 71 72 72 72 72 72 72 72 73 73 73
## [51] 73 73 73 73 74 74 74 75
```

```r
print(paste("Number of datapoints:",length(pr97to99)))
```

```
## [1] "Number of datapoints: 58"
```

```r
print(paste("Median College_Score:",median(pr97to99)))
```

```
## [1] "Median College_Score: 70"
```

```r
print(paste("Mean College_Score:",round(mean(pr97to99), digits=1)))
```

```
## [1] "Mean College_Score: 68"
```

Given the author's **HighSchool_PR** 95 score, **College_Score** 69 is an above-average outcome. Although it is somewhat unfair to compare **HighSchool_PR** 95 with **HighSchool_PR** 97 as a nearby value, **College_Score** 69 is only one point below the median **College_Score** of the **HighSchool_PR** 97-99 group. We do not know whether the positive outcome is due to the fact that the author studied extra hard given the more competitive environment in her high school, and/or she received more resources than most respondents with **HighSchool_PR** 95.

The most prestigious high schools in Taipei typically requires **HighSchool_PR** 97 or higher.[119] Students in Taipei with **HighSchool_PR** 97 may receive significantly more resources than the ones with **HighSchool_PR** 95, despite the difference is only two percentage points. This phenomenon is less likely in areas where the **HighSchool_PR** requirement is not as high in their top local high schools.

## 13.2 Comparison with the Model Prediction

We also apply the logistic regression model in Section 7.2, which predicts the probability of a respondent achieving **College_Score** at least 65 with his/her **HighSchool_PR**. Given the author's **HighSchool_PR** 95 score, the predictive probability is about 64.6% in terms of how likely she would have achieved **College_Score** at least 65. As a reference, Section 7.3 shows that given **HighSchool_PR** 99, the estimated probability to achieve this is 76.9%. The numbers made the author feel better, because she still had a good chance of getting **College_Score** at least 65 given her original **HighSchool_PR**.

```r
original_model = glm(CS_65up ~ HighSchool_PR, data=data_corr, family="binomial")
# summary(original_model)

my_data = data.frame(HighSchool_PR = 95, College_Score = 69, CS_65up = TRUE)

pred_prob = predict.glm(original_model, my_data, type="response")
# type="response" gives the predicted probabilities

pred_prob
```

```
##         1
## 0.6464676
```

Again, the logistic regression model assumes all else are held equal. We also assume that most respondents attended a high school whose admission requirement is about their own **HighSchool_PR**. From this data, we cannot measure the effect from impostor syndrome – one decided to study extra hard to compensate that

[119] https://cclccl-life.blogspot.com/2013/06/blog-post_9.html

he/she had a much lower **HighSchool_PR** compared with his/her classmates in high school. Given the goal of demonstrating statistical methods in this project, we do not think it is practical to find such people and compare their scores with the people who did not receive this opportunity. This can be a full research project in the education field.

# Final Words

The author has been building this project on and off since 2019. With limited time outside her full-time job and various hobbies, she persevered and completed this whole document little by little. Depending on her availability, some days she wrote multiple paragraphs in a single commit, while other days she wrote as few as 2-3 sentences. Writing a large document may seem challenging in the beginning, but the author demonstrated that it is possible to finish one through accumulation of efforts. There is no secret ingredient – just keep writing. Here is a quote from Jodi Picoult, an American bestselling author:[120] "You can always edit a bad page. You can't edit a blank page."[121] You do not have to write every day, but you have to be consistent in the process. It is also acceptable to pause for a few weeks (or even months), and come back to the project again.

In addition to the motivations stated in the Introduction (Chapter 1) and the Personal Scores and Remarks (Chapter 13), the author was also inspired by the book *Build a Career in Data Science* (Robinson and Nolis, 2020). She saw some interesting meta-analysis about the writing process of the book,[122] such as how many words were written by which person and at what time of the day. She realized that writing a book and/or a long manuscript takes tremendous work,[123] but the knowledge-sharing advantages are well worth the time and energy spent.[124] Publishing something in the open world not only builds your portfolio, but also increases your visibility and exposure. You are also likely to receive a few positive responses from readers, including potential collaboration invitations.

This may sound like a cliche, but the author definitely learned a lot throughout the project – not just the statistical content. Hard skills include and are not limited to Git, R, and writing reproducible code. One of the author's favorite part is generating "green squares" (GitHub commits) as tangible evidence of progress.[125] It feels like planting virtual trees on her own GitHub garden. The author also gained soft skills from the process, such as clear communication in writing and time management. Most importantly, she learned how to break down a large project into smaller tasks. In data science job interviews, she encountered some questions relevant to concepts in this project, such as the conditional probability and model metrics. Side benefits of writing this document are to refresh your knowledge and deepen your understanding of data science. We never know when the skills will come in handy in the future.

This project is the longest manuscript the author has independently completed after her PhD dissertation (Chai, 2017). Other long-form publications she has written as a single author include her review article on text preprocessing (Chai, 2023b) and her Masters thesis (Chai, 2013). The author has collaborated with others on a wide range of topics throughout the years, including Bayesian statistics (Lu and Chai, 2022; Clyde et al., 2018), federal survey analysis (Avery et al., 2017), and other statistical applications (Henry et al., 2019; Beckman et al., 2015). The author is also a regular contributor to the Royal Statistical Society (RSS)[126] discussion papers, in which she writes constructive feedback up to 400 words per session. Examples are (Chai, 2024), (Chai, 2023a), and (Chai, 2021).

[120] https://www.jodipicoult.com/
[121] https://www.writingforward.com/better-writing/you-cant-edit-a-blank-page
[122] https://jnolis.com/blog/data_science_on_book/
[123] https://hookedondata.org/posts/2021-04-06_publishing-a-technical-book-part-3/
[124] https://hookedondata.org/posts/2021-04-06_publishing-a-technical-book-part-1/
[125] https://www.palomamedina.com/biceps
[126] https://rss.org.uk/

## Acknowledgments

The author would like to thank Dr. Mine Cetinkaya-Rundel and Dr. David Banks at Duke University; they both motivated the author to teach statistics and create reproducible work in R. The author is also grateful to her former Microsoft colleagues Smit Patel and Dylan Stout for troubleshooting GitHub issues.

The author would also like to acknowledge Dr. Cliburn Chan and Dr. Janice McCarthy for introducing her to GitHub in the statistical computation course at Duke University. This provided her the foundations to use GitHub as a modern version control system in the first place.

Finally, the author gives a special mention to her significant other, Hugh Hendrickson, for all his support in the author's professional career development.

## References

Agresti, A. (2003). *Categorical Data Analysis*. John Wiley & Sons, Hoboken NJ, United States.

Allaire, J., Horner, J., Xie, Y., Marti, V., and Porte, N. (2019). *`markdown`: Render Markdown with the C Library 'Sundown'*. R package version 1.1.

Alpaydin, E. (2020). *Introduction to Machine Learning*. MIT Press, Cambridge MA, United States, 4th edition. MIT stands for Massachusetts Institute of Technology.

American Statistical Association (2016). Statement on statistical significance and p-values. *The American Statistician*, 70(2):129–133.

Avery, R. B., Bilinski, M. F., Bucks, B. K., Chai, C., Chow, M., Clement, A., Critchfield, T., Frumkin, S., Keith, I. H., Mohamed, I. E., Pafenberg, F. W., Patrabansh, S., Schultz, J. D., and Wood, C. E. (2017). A profile of 2014 mortgage borrowers: Statistics from the National Survey of Mortgage Originations. Technical report, National Mortgage Database, Federal Housing Finance Agency, Washington DC, United States.

Baumer, B., Cetinkaya-Rundel, M., Bray, A., Loi, L., and Horton, N. J. (2014). R Markdown: Integrating a reproducible analysis tool into introductory statistics. *arXiv preprint arXiv:1402.1894*.

Beckman, E., Chai, C., Lyu, J., Mahserejian, S., Tran, H., Yavari, S., Mitchell, H., Calatroni, A., and Kang, E. L. (2015). Investigating the relationship between the microbiome and environmental characteristics. In *Twentyfirst Mathematical and Statistical Modeling Workshop for Graduate Students*, pages 89–112. North Carolina State University.

Bertin, J. (1983). *Semiology of Graphics: Diagrams, Networks, Maps*. The University of Wisconsin Press, Madison WI, United States. Originally published in French. Translated to English by William J. Berg in 2010.

Casella, G. and Berger, R. L. (2021). *Statistical Inference*. Cengage Learning, Boston MA, United States.

Chai, C. P. (2013). Facebook account misuse detection – A statistical approach. Master's thesis, National Taiwan University, Taipei, Taiwan.

Chai, C. P. (2017). *Statistical Issues in Quantifying Text Mining Performance*. PhD dissertation, Duke University, Durham NC, United States.

Chai, C. P. (2020). The importance of data cleaning: Three visualization examples. *CHANCE*, 33(1):4–9. Available from: https://chance.amstat.org/2020/02/data-cleaning/.

Chai, C. P. (2021). Christine P. Chai's contribution to the discussion of "Testing by betting: A strategy for statistical and scientific communication" by Glenn Shafer. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 184(2):449–450.

Chai, C. P. (2023a). Christine P. Chai's contribution to the discussion of "The first discussion meeting on statistical aspects of climate change". *Journal of the Royal Statistical Society Series C: Applied Statistics*, 72(4):854–855.

Chai, C. P. (2023b). Comparison of text preprocessing methods. *Natural Language Engineering*, 29(3):509–553. Available from: https://doi.org/10.1017/S1351324922000213.

Chai, C. P. (2024). Christine P. Chai's contribution to the discussion of "A system of population estimates compiled from administrative data only" by Dunne and Zhang. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 187(1):31–32.

Chen, H.-L. S. and Huang, H.-Y. (2017). Advancing 21st century competencies in Taiwan. Technical report, Asia Society – Center for Global Education, New York NY, United States.

Chen, P. (2008). Strategic leadership and school reform in Taiwan. *School Effectiveness and School Improvement*, 19(3):293–318.

Chiang, Y.-L. (2022). Exams or applications? Elite Taiwanese students' perceptions and navigation of college admissions systems. *International Journal of Comparative Sociology*, 63(1-2):30–50.

Chou, C. P. (2015). Higher education development in Taiwan. In *Mass Higher Education Development in East Asia*, pages 89–103. Springer, Cham, Switzerland.

Chou, C. P. and Ho, A.-H. (2007). Schooling in Taiwan. In *Going to School in East Asia*, pages 344–377. Greenwood, New York NY, United States.

Clyde, M., Cetinkaya-Rundel, M., Rundel, C., Banks, D., Chai, C., and Huang, L. (2018). *An Introduction to Bayesian Thinking*. GitHub, San Francisco CA, United States, 1st edition. Available from: https://statswithr.github.io/book/.

DeGroot, M. H. and Schervish, M. J. (2012). *Probability and Statistics*. Pearson Education, Boston MA, United States, 4th edition. Available from: http://bio5495.wustl.edu/Probability/Readings/DeGroot4thEdition.pdf.

Diez, D. M., Cetinkaya-Rundel, M., and Barr, C. D. (2019). *OpenIntro Statistics*. OpenIntro, Boston MA, United States, 4th edition. Available from: https://www.openintro.org/book/os/.

Dusetzina, S. B., Tyree, S., Meyer, A.-M., Meyer, A., Green, L., and Carpenter, W. R. (2014). An overview of record linkage methods. In *Linking Data for Health Services Research: A Framework and Instructional Guide [Internet]*, chapter 4, pages 29–49. Agency for Healthcare Research and Quality (US), Rockville MD, United States. Available from: https://www.ncbi.nlm.nih.gov/sites/books/NBK253313/pdf/Bookshelf_NBK253313.pdf.

Foote, S. (2014). *Learning to Program*. Addison-Wesley Professional, Boston MA, United States.

Gilleland, E. (2015). *`verification`: Weather Forecast Verification Utilities*. R package version 1.42.

Goodman, S. (2008). A dirty dozen: Twelve p-value misconceptions. In *Seminars in Hematology*, volume 45, pages 135–140. Elsevier, Amsterdam, Netherlands.

Han, H. (2022). The utility of receiver operating characteristic curve in educational assessment: Performance prediction. *Mathematics*, 10(9):1493.

Henry, T. R., Banks, D., Owens-Oas, D., and Chai, C. (2019). Modeling community structure and topics in dynamic text networks. *Journal of Classification*, 36(2):322–349.

Hoff, P. D. (2009). *A First Course in Bayesian Statistical Methods*. Springer, New York NY, United States.

Hsieh, T.-L. (2019). A preliminary study of multiple college admission criteria in Taiwan: The relationship among motivation, standardized tests, high school achievements, and college success. *Higher Education Research & Development*, 38(4):762–779.

Kanter, J. M. and Veeramachaneni, K. (2015). Deep feature synthesis: Towards automating data science endeavors. In *2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pages 1–10. IEEE.

Kruschke, J. K. and Liddell, T. M. (2018). The Bayesian new statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. *Psychonomic Bulletin & Review*, 25(1):178–206.

Liu, R.-F. (2022). Pathways to college admissions: Student strategies and class variations in activating cultural knowledge in Taiwan. *International Studies in Sociology of Education*, 31(3):284–304.

Lock, R. H., Lock, P. F., Morgan, K. L., Lock, E. F., and Lock, D. F. (2020). *Statistics: Unlocking the Power of Data*. John Wiley & Sons, Hoboken NJ, United States, 3rd edition. Available from: https://www.lock5stat.com/.

Lu, J. and Chai, C. P. (2022). Robust Bayesian nonnegative matrix factorization with implicit regularizers. *arXiv preprint arXiv:2208.10053*.

Martin, R. C. (2009). *Clean Code: A Handbook of Agile Software Craftsmanship*. Pearson Education, Boston MA, United States, 1st edition. Available from: https://tinyurl.com/clean-code-pdf.

McConnell, S. (2004). *Code Complete*. Pearson Education, Boston MA, United States, 2nd edition. Available from: https://tinyurl.com/code-complete-book.

NCSS (n.d.). Binary diagnostic tests – single sample. In *NCSS Statistical Software*, chapter 535. NCSS (Number Cruncher Statistical System), Kaysville UT, United States. Available from: https://ncss-wpengine.netdna-ssl.com/wp-content/themes/ncss/pdf/Procedures/NCSS/Binary_Diagnostic_Tests-Single_Sample.pdf.

Ramsey, F. and Schafer, D. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis*. Cengage Learning, Boston MA, United States. Data files are available on: http://www.statisticalsleuth.com/.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., and Müller, M. (2011). `pROC`: An open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, 12:77.

Robinson, E. and Nolis, J. (2020). *Build a Career in Data Science*. Manning Publications, Shelter Island NY, United States. Available from: https://www.manning.com/books/build-a-career-in-data-science.

Sheather, S. (2009). *A Modern Approach to Regression with* R. Springer, New York NY, United States.

Shy, H.-Y., Chiu, C.-Y., Chiang, M.-W., and Liao, S.-C. (2021). Comparison of time management ability between medical students who entered medical universities through different approaches. *ResearchSquare preprint rs-642312/v1*. Available from: https://doi.org/10.21203/rs.3.rs-642312/v1.

Wasserstein, R. L., Schirm, A. L., and Lazar, N. A. (2019). Moving to a world beyond "p < 0.05". *The American Statistician*, 73(sup1):1–19.

Wickham, H. (2016). *`ggplot2`: Elegant Graphics for Data Analysis*. Springer-Verlag New York, New York NY, United States. Available from: https://ggplot2.tidyverse.org.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K., and Yutani, H. (2019). Welcome to the `tidyverse`. *Journal of Open Source Software*, 4(43):1686.

Wickham, H. and Grolemund, G. (2016). R *for Data Science: Import, Tidy, Transform, Visualize, and Model Data*. O'Reilly Media, Inc., Sebastopol CA, United States. Available from: https://www.tidyverse.org/.

Wilkinson, L. (2013). *The Grammar of Graphics*. Springer, New York NY, United States, 2nd edition.

Xie, Y. (2016). *`bookdown`: Authoring Books and Technical Documents with* R *Markdown*. Chapman and Hall/CRC, Boca Raton FL, United States. R package version 0.25. Available from: https://bookdown.org/yihui/bookdown.

Yang, S. and Berdine, G. (2017). The receiver operating characteristic (ROC) curve. *The Southwest Respiratory and Critical Care Chronicles*, 5(19):34–36.

Yau, N. (2011). *Visualize This: The FlowingData Guide to Design, Visualization, and Statistics*. John Wiley & Sons, Hoboken NJ, United States.