PTT Analysis of Entrance Exam Scores in Taiwan

Christine P. Chai

February 21, 2021

Ongoing work since 2019.

Executive Summary

Write something here

1   Introduction

The author(s) 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, so some people studied little in middle school and was able to get into a good high school. Then some of these people kept studying little and ended up with a bad score on the college entrance exam. On the other hand, we have also seen some students from mediocre high schools studied very hard, and eventually earned a stellar score in the college exam. Therefore, we decided to gather data and create our own analysis. The target audience of this document would be people with a basic understanding of statistics, equivalent to taking or having completed Statistics 101.

Since we do not have the official data from the Taiwan government that links people between their scores on the high school entrance exam and college entrance exam, it is difficult to obtain a representative sample of people's scores. Therefore, we obtained self-reported data from an existing post[2] 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.

For the rest of this article, we begin with a short background of Taiwan's high school and college entrance exams, followed by the description of our dataset. We start the technical content with exploratory data analysis, in order to identify and visualize the distribution of exam scores. Next, we explain our decisions in statistical modeling, such as whether a linear regression is appropriate or not. We also take a closer look at the top scorers, because they typically came from prestigious high schools and/or colleges. Then we move on to introduce advanced models like logistic regression. Finally, we provide our discussion, state our conclusion, and give some personal remarks.

2   Background

In Taiwan, the high school entrance exam score percentiles are between 1% and 99%, and people often refer to the percentile rank (PR value) as from 1 to 99. This scoring system existed from 2001 to 2013[3]. The actual exam score ranges were different. For example, the maximum possible score was originally set to 300, but it was increased to 312 in Year 2007. Then the maximum possible score was increased to 412 in Year 2009.

---

[1]https://bit.ly/2JSPXKc
[2]https://www.ptt.cc/bbs/SENIORHIGH/M.1432729401.A.995.html
[3]https://bit.ly/2JNQaOI

Therefore, the percentile ranks (PR values) serves as a normalized tool to compare academic achievements across different years.

The college entrance exams are held twice a year in Taiwan; the first exam is for early admission and the second exam is for regular admission. The first exam, typically held in late January or early February, is called the General Scholastic Ability Test (GSAT)[4]. The second exam is called the Advanced Subjects Test (AST)[5], and it is almost always held on July 1st, 2nd, and 3rd. The GSAT scores are normalized to a range of 0 to 75, regardless of the difficulty level of GSAT each year. On the other hand, the scores of AST can vary widely because each subject is scored separately from 0 to 100. Since the AST scores fluctuate more due to the difficulty level of the exam questions each year, we decided to use the GSAT scores as a benchmark of the college exam scores.

Remark: The GSAT consists of five subjects, each of which are graded on a 0 to 15 point scale. Starting in 2019, students may choose four of the five subjects for the GSAT. The maximum possible score (i.e., full marks) is reduced from 75 to 60.

3 Data Description

It is a challenge to obtain individual pairs of data as a representative sample. Although it is easy to send out a spreadsheet and ask 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.

Hence we retrieved data from the SENIORHIGH (high school)[6] discussion section on PTT[7], the largest terminal-based bulletin board in Taiwan.[8] We assume the data to be more representative (than if we had collected on our own) because anyone could get a PTT account and reply to the post – at least before the restriction was announced in 2018.[9] The majority of scores were reported in May 2015, and a few scores were reported in the following month or later.

The data `ptt_SENIORHIGH_data.csv` contain 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 we recorded each of them as "-1" (an invalid numerical value).

In some cases, the reported scores can be inaccurate based on the respondent's description, so we created two indicators for this issue:

- **HS_Inacc**: A "1" means the reported score of high school entrance exam is inaccurate.
- **College_Inacc**: A "1" means the reported score of college entrance exam is inaccurate.

Some people reported their percentile rank (PR) from the mock exam, rather than 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.[10]

---

[4]https://bit.ly/2W0fdUq
[5]https://bit.ly/2J7YxoW
[6]https://www.ptt.cc/bbs/SENIORHIGH/M.1432729401.A.995.html
[7]If you have a PTT account, you can log into the website using a browser. https://iamchucky.github.io/PttChrome/?site=ptt.cc
[8]https://en.wikipedia.org/wiki/PTT_Bulletin_Board_System
[9]https://www.ettoday.net/news/20200304/1659455.htm
[10]https://tsjh301.blogspot.com/2014/06/compulsory-education.html

Moreover, 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.

3.1 Data at a Glance

We show the first 10 rows of data here, and `NA` (not available) denotes that the value is missing. Note that only **HS_Inacc** and **College_Inacc** contain `NA`s, because we already recoded missing values to "-1" (an invalid numeric value) for **HighSchool_PR** and **College_Score**.

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

data[1:10,]
```

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

Remark: 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.[11] 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 [3]. Our dataset `ptt_SENIORHIGH_data.csv` is relatively clean, but we still had to recode and flag missing values.

4 Exploratory Data Analysis

The first step in a data project is exploratory data analysis, before we perform any statistical modeling. Therefore, we start with observing the trends of the two main variables, **HighSchool_PR** and **College_Score**.

4.1 High School Entrance Exam Scores (Percentile Rank)

Below shows the descriptive statistics of **HighSchool_PR**, i.e., the percentile rank of high school entrance exam scores. The missing values are removed beforehand. Approximately 75% of the respondents have a percentile rank (PR) at least 85, indicating that most of the respondents scored in the top 15% of the high school entrance exam. The histogram is also extremely left-skewed.
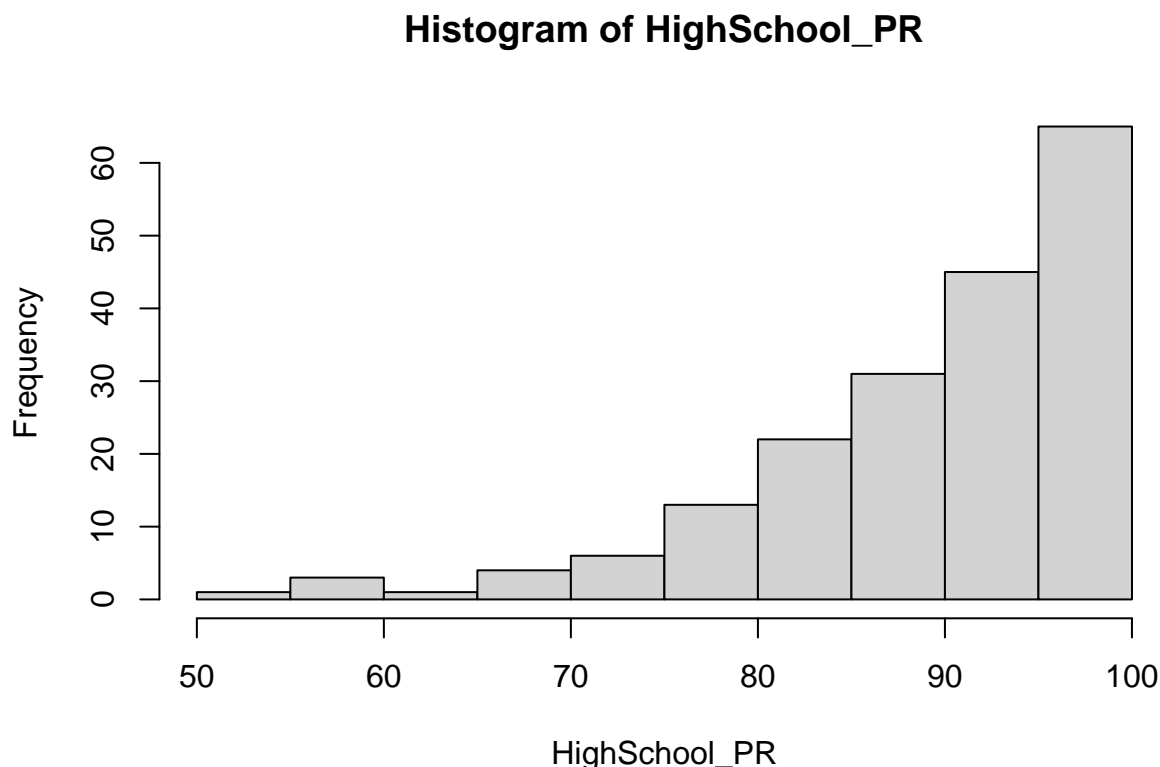
```
# High school entrance exam scores: Remove missing values
uni_HS_score = data$HighSchool_PR[which(data$HighSchool_PR != -1)]

summary(uni_HS_score)
```

---

[11]https://bit.ly/303IWxY

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

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

Histogram of 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 0 and 75. The histogram is also left-skewed, but less extreme than **HighSchool_PR**.

According to the reference score table[12] 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 did not limit registration to students of these colleges in the past, so the population of PTT is slightly more diverse. Note that as of 2020, PTT changed their eligibility requirements, and limited new account registrations to only people with an email address from National Taiwan University.[13]

---

[12]https://bit.ly/3bAYOvO
[13]Screenshot obtained on May 26, 2020: https://imgur.com/33fwrGH

```
# College entrance exam scores: Remove missing values
uni_college_score = data$College_Score[which(data$College_Score != -1)]

summary(uni_college_score)
```

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

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

Histogram of College_Score
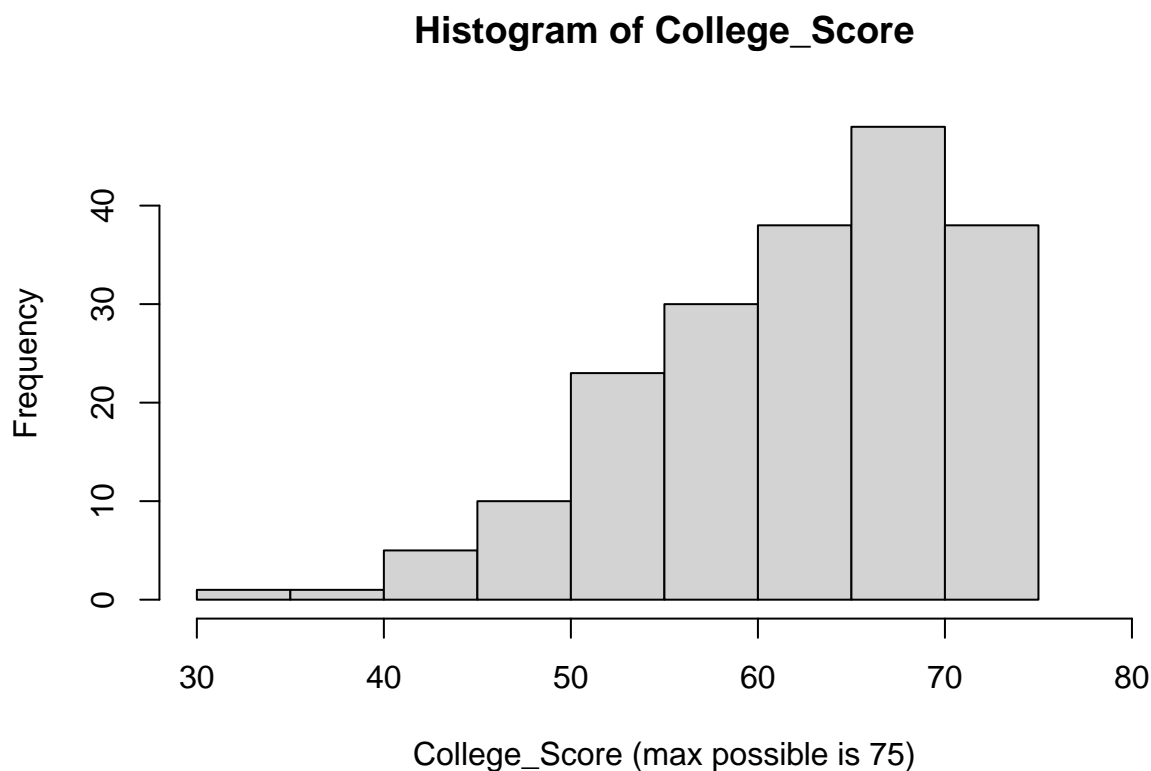
4.3 Bivariate Exploration

Next, we create a bivariate scatterplot of **HighSchool_PR** and **College_Score**, but we have to remove the records with at least one missing score. Just like what we observed in the univariate plots, both variables are largely concentrated towards the maximum possible scores.

```
missing_rows = which(data$HighSchool_PR == "-1" | data$College_Score == "-1")
# Indices: 6  19  71  85  88  96 132 183 195 => nine in total

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

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

High School and College Entrance Exam Scores

The correlation coefficient is approximately 0.507, showing a medium strength of positive association between **HighSchool_PR** and **College_Score**. We can interpret that a better score in the high school entrance exam is likely to lead to a better college entrance exam score, but the relationship is not as strong after **HighSchool_PR** reaches 80.

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

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

To calculate the correlation coefficient between the random variables X, Y, we need to 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 σ_X:

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

Same applies to the standard deviation σ_Y.

Finally, we can calculate the correlation coefficient as:

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

5 Linear Regression

We are going to decide whether we should run a linear regression to predict **College_Score** from **High-School_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 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. α is the intercept, and it stands for the estimate $\hat{Y}$ when $X = 0$ (if applicable). β is the coefficient, and when X increases by one unit, we can expect Y to increase by β units. Last but not least, ϵ is the error term, which is normally distributed with mean zero.

OpenIntro Statistics [5] states that four requirements need to be met for a 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")
```

High School and College Entrance Exam Scores


```r
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.

```r
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)
```

High School and College Entrance Exam Scores

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** ≥ 80 and **College_Score** ≥ 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** ≥ 80 / all respondents.

- $P(\text{College\_60up is true}) = \#$ of respondents with **College_Score** ≥ 60 / all respondents.

- $P(\text{HighSchool\_80up is true} \cap \text{College\_60up is true})$
  $= \#$ of respondents with **HighSchool_PR** ≥ 80 and **College_Score** ≥ 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.

$$P(\text{College\_60up is true}) = \text{\# of respondents with College\_Score} \geq 60 \text{ / all respondents}$$
$$= \frac{120}{188} \approx 0.6382.$$

```
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: A Companion to the Statistics with* R *Course* [4] as a starting point to learn Bayesian statistics. It is the supplementary materials for the Bayesian statistics course on Coursera from Duke University[14].

6 Top Scorers: A Closer Look

We are going to take a closer look at the top scorers, given the observation 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 top tier can be very different than the bottom of top tier. Therefore, we consider the following subcategories:

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

6.1 Data Consistency

Before we start the analysis, we need to ensure consistency in the data. For instance, there are 191 records of valid high school entrance exam scores in the data. But if we consider only the ones with a valid college entrance exam 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 is going to be problematic when we combine the univariate analysis with the bivariate analysis. Thus, we should use only the 188 records whose college entrance exam scores are also valid.

```
# High school entrance exam scores: Remove missing values
# uni_HS_score = data$HighSchool_PR[which(data$HighSchool_PR != -1)]
length(uni_HS_score)
```

[1] 191

```
# Consider only the records with both valid HighSchool_PR and College_Score
length(data_corr$HighSchool_PR)
```

[1] 188

The same data consistency requirement also applies to the college entrance scores. There are 194 records of valid high school entrance exam scores in the data, but only 188 of them also have corresponding valid high school entrance exam scores. Accordingly, we should use only the 188 records whose high school entrance exam scores are also valid.

```
# College entrance exam scores: Remove missing values
# uni_college_score = data$College_Score[which(data$College_Score != -1)]
length(uni_college_score)
```

[1] 194

---

[14]https://www.coursera.org/learn/bayesian

```
# Consider only the records with both valid HighSchool_PR and College_Score
length(data_corr$College_Score)
```

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

6.2 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")
```

Histogram of HighSchool_PR 80 and above

12

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[15]. 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[16].

```
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"
```

6.3 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[17]. 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)
```

---

[15]https://w199584.pixnet.net/blog/post/28321887

[16]http://www.edtung.com/TopNews/NewsContent.aspx?type=4&no=1278

[17]https://web.archive.org/web/20150207042900/http://www.ceec.edu.tw/AbilityExam/AbilityExamStat.htm

```
data.frame(college_score, college_pr)
```

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


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"
```

6.4   Bivariate Exploration of High 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** $\geq$ 80 and/or **College_Score** $\geq$ 60. There are 163 respondents with **HighSchool_PR** $\geq$ 80, 128 respondents with **College_Score** $\geq$ 60, and 120 respondents with both. Since the number of respondents with **HighSchool_PR** $\geq$ 80 does not equal to the number of respondents with **College_Score** $\geq$ 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.

```
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 shows a positive association between **HighSchool_PR** and **College_Score**: 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 findings 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**.

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 Section 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. α is the intercept, and β is the coefficient. When β increases by one unit, the log odds $\log(\frac{p}{1-p})$ increases by β units. In other words, when β 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 α 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* [1] 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.

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

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

```
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.

```
# 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)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7122  -0.9948  -0.1334   0.8248   2.5246
##
## 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

18

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.

```
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.

```
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.

```
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.)

```
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.

```
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 shows 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) has 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.

$$P(\textbf{HighSchool\_PR} = 99|\textbf{College\_Score} \geq 65) = \frac{24}{92} \approx 26\%.$$

```
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)

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

```
## [1] "Number of respondents with HighSchool_PR 99: 25"
```
```
print(paste("Number of respondents with College_Score 65 or better:",num_cs65))
```

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

```
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. 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.

```
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
```

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 0.769102549200389"
```

```
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")
```

Predicted Probability of College_Score >= 65 with HighSchool_PR

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")
```

Jittered Probability of College_Score >= 65 with HighSchool_PR

8 Model Validation

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* [2]. 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.

However, 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. In-sample prediction is insufficient because we need to test on unseen data to avoid overfitting.[18] 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. Thus, we need to perform out-of-sample prediction on the model; that is, train the model on some data and test it on different data.

In the later sections, we will demonstrate two methods to implement out-of-sample prediction, and the

---

[18]https://elitedatascience.com/overfitting-in-machine-learning

first method is to **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. 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.[19] Finally we test 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.

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. One common implementation is the **K-fold cross validation**, in which 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. Another common implementation is the **leave-one-out cross validation**, in which 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, but we need to perform N rounds of validation, which may not be feasible when N is extremely large. We will demonstrate both K-fold cross validation and leave-one-out cross validation in this chapter.

8.1 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.

---

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

```
# 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[20]:

- **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.

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

---

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

```
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.

```
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.

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.[21] 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,

---

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

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**.

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

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.

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

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 [7], 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.[22] 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.[23]

---

[22]https://fcase.io/a-major-challenge-false-positives/

[23]https://brownmath.com/stat/falsepos.htm

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 importance 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.[24]

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.

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

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.

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

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

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

Let's see an example in medical testing.[25] 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.

---

[24]https://www.healthnewsreview.org/toolkit/tips-for-understanding-studies/understanding-medical-tests-sensitivity-specificity-and-positive-predictive-value/

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

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.[26] 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.

Unfinished below

8.3   Breakdown by High School Entrance Exam Scores

Before starting this section, complete the foreword for the Model Validation chapter (done).

Examine the confusion matrices for each group of **HighSchool_PR**: below 80, 80-89, 90-94, 95-99.

Ensure that each group has a sufficiently large number of respondents.

Refer to Section 6.2 for the details of this categorization.

Create more graphs

8.4   Out-of-Sample Prediction

Demonstrate both K-fold cross validation and leave-one-out cross validation!

The goal of writing this document is to instruct the readers on how to perform data analysis using fundamental machine learning concepts.

Focus on demonstrating the methodology

Explain why overfitting is bad!

Randomly divide the data into a training set and a testing set
Set a random seed to ensure reproducibility

If fine-tuning the model is needed: Training set, validation set, testing set
Avoid overfitting the testing set!

Consider cross validation (put this topic after the training-testing method)

Draft code for training and testing partitions

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. 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)
```

---

[26]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
train_inds = ind_permute[1:94]
test_inds = ind_permute[95:188]

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

Unfinished below

Print the train_inds and test_inds

```r
# Draft code for training and testing partitions

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

# train_model

test_prob = predict.glm(train_model, test_data, type="response")
# type="response" gives the predicted probabilities

# ---------------------------------

# 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
```

```r
# percentage version
prop.table(test_confusion)
```

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

```r
# UNFINISHED: Organize the initial code and start writing the descriptions
# Also consider making part of it into a function ??
```

Draft code for k-fold cross validation

```r
# Draft code for k-fold cross validation

set.seed(21) # different seed than the training-testing partitions

nn = nrow(data_corr) # 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

# 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)

partition_list = list(0,0,0,0,0,0,0,0,0,0)
partition_list[[1]] = 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]] = ind_permute[start:end]
}

# partition_list

# UNFINISHED: Use the partitions for training and testing
```

9 Discussion and Conclusion

The Statistics 101 course provides a starting point for students to perform data analysis. Linear regression is widely used, but it is not a panacea for data analysis. The model assumptions need to be met in the data, as stated at the beginning of Section 5.

For the next steps in learning statistics, we recommend reading *The Statistical Sleuth: A Course in Methods of Data Analysis* [8], which is the textbook for undergraduate-level regression analysis at Duke Statistical Science.[27] 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.[28]

Write something more

10 Final: Personal Remarks

Write something here

Taipei First Girls' High School[29] typically requires **HighSchool_PR** 99 for admission. There are some exceptions, such as recruited athletes, students with disabilities,[30] and students under other extraordinary situations.[31]

The Department of Electrical Engineering at National Taiwan University (NTUEE)[32] typically requires full

---

[27]https://www2.stat.duke.edu/courses/Fall18/sta210.001/
[28]http://www.statisticalsleuth.com/
[29]http://web.fg.tp.edu.tw/~tfghweb/EnglishPage/index.php
[30]http://www.rootlaw.com.tw/LawArticle.aspx?LawID=A040080080001900-1020822
[31]https://bit.ly/2WtRY63
[32]https://web.ee.ntu.edu.tw/eng/index.php

marks (15 out of 15) in English, mathematics, and science in **College_Score** for the early admission.[33] 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 regular college entrance exam process in July.

Finally, include my own scores as a datapoint for prediction.

Don't show the numbers until I am ready to work on this section!

Acknowledgments

The author would like to thank her Microsoft colleagues Smit Patel and Dylan Stout for troubleshooting GitHub issues.

More to add

Appendix

RStudio, code reproducibility, etc.

Need to specify the R libraries.

The code is mainly for demonstration purposes. To learn about better coding styles, we recommend the book *Code Complete* [6] instead.

References

[1] Alan Agresti. *Categorical Data Analysis*, volume 482. John Wiley & Sons, 2003. Available from: https://tinyurl.com/categorical-data-analysis.

[2] Ethem Alpaydin. *Introduction to Machine Learning*. MIT press, 4th edition, 2020.

[3] Christine P Chai. The importance of data cleaning: Three visualization examples. *CHANCE*, 33(1):4–9, 2020. Available from: https://chance.amstat.org/2020/02/data-cleaning/.

[4] Merlise Clyde, Mine Cetinkaya-Rundel, Colin Rundel, David Banks, Christine Chai, and Lizzy Huang. *An Introduction to Bayesian Thinking: A Companion to the Statistics with* R *Course*. GitHub, 2020. Available from: https://statswithr.github.io/book/.

[5] David M Diez, Mine Cetinkaya-Rundel, and Christopher D Barr. *OpenIntro Statistics*. OpenIntro, 4th edition, 2019. Available from: https://www.openintro.org/book/os/.

[6] Steve McConnell. *Code Complete*. Pearson Education, 2nd edition, 2004. Available from: https://tinyurl.com/code-complete-book.

[7] NCSS. Binary diagnostic tests – single sample. In *NCSS Statistical Software*, chapter 535. NCSS (Number Cruncher Statistical System), n.d. Available from: https://ncss-wpengine.netdna-ssl.com/wp-content/themes/ncss/pdf/Procedures/NCSS/Binary_Diagnostic_Tests-Single_Sample.pdf.

[8] Fred Ramsey and Daniel Schafer. *The Statistical Sleuth: A Course in Methods of Data Analysis*. Cengage Learning, 2013. Data files are available on: http://www.statisticalsleuth.com/.

---

[33]https://university.1111.com.tw/univ_depinfo9.aspx?sno=100102&mno=520101