ST447 Data Analysis and Statistical Methods

Passing rate prediction for Driving Test in the UK

Candidate No – 50175

MSc Data Science

Introduction

Our problem description defines a LSE student, XYZ that wants to give the driving test at one of two centres;

the test centre nearest to his/her home or the test centre nearest to LSE.

In this paper, we will intend to help XYZ make this decision, by analysing the dataset DVSA1203 is available at https://www.gov.uk/government/statistical-data-sets/car-driving-test-data-by-test-centre which contains information on car pass rates by age (17 to 25 year olds), gender, year (2007-2022) and test centre.
This project will implement a logistic regression model and other parametric/non-parametric tests in R, for our student XYZ who is a male aged 25, living in Nottingham (Colwick).

Student	Age	Gender	Address
XYZ	25	Male	Nottingham (Colwick)

The test centre nearest to home is the **Nottingham (Colwick)** centre and the **Wood Green (London)** centre is closest to LSE.

A **Logistic Regression Model** is quite similar to a Linear Regression, but it is used for binary classification.

The logistic function can be formulated as

$$\text{logistic}(\eta) = \frac{1}{1 + \exp(-\eta)}$$

For classification, we prefer probabilities between 0 and 1, so we wrap the right side of the linear regression equation into the logistic function. This forces the output to assume only values between 0 and 1.

$$P(y_{(i)} = 1) = \frac{1}{1 + \exp(-(\beta_0 + \beta_1 x_1^{(i)} + \cdots + \beta_p x_p^{(i)})}$$

We shall implement this model and make predictions on its basis.

Methodology

The entire data for all the years has been cleaned and processed in R to get an understanding of the Passing Rates in all the cities over the years. We try to establish a known population distribution for the entire dataset, using plots like qqplots and tests like the **Kolmogorov- Smirnov** test.

We will try to infer the possibility of a trend for the Passing rate over the years, and patterns that can help us make a better prediction for our student. Realising the stochastic nature of the yearly Passing rates, we will then create a logistic regression model using **all the data** available to us.

We will take the variables Year, Age, and the Location.

- Since we are only interested in the two centres Nottingham (Colwick) and Wood Green (London); we will one-hot encode the locations as binary responses where Wood Green centre will be 0 and Nottingham centre will be 1.
- We will take the variable Age as it is.
- Due to the year being a factor variable in this case, we will encode Year as a vector of 1 to 15, where 2021-22 is represented by 1 and 2007-08 by 15.

We don't consider Gender as a variable, as we our concerned with only the Male passing rate data. Hence, we take the assumption that adding a binary categorical variable of gender, which may be significant, won't help us in emphasizing the difference between the Passing rates in the two centres.

R Code

Profile

```
ID = 202249724
source('XYZprofile.r')
XYZprofile(ID)

## The profile of XYZ:
## - Age:  25
## - Gender:  Male
## - Home address:  Nottingham (Colwick)
```

Importing libraries

```
library(readODS)
library(dplyr)
library(ggpubr)
library(ggplot2)
library(tidyverse)
library(CatEncoders)
```

Cleaning and wrangling the data

```
datall = read.ods(file = 'dvsa1203.ods')
#This function reads all the sheets of the ods file and returns them as datafram
elements of a list,(16 elements for 15 sheets)

dat = datall[2:16] #The first sheet contains content and metadata.
```

Looping through our list of dataframe,

```
cleaned_data = list()
for (i in 1:length(dat)) {
    cleaned_data[[i]] <- dat[[i]][-(1:7),]
    #Removing metadata; {first 6 rows and the column names(we will assign new
```

```
names later)}
    if (i > 7){#Removing empty columns in some sheets & fixing their column name
        cleaned_data[[i]] <- cleaned_data[[i]][-c(6,10)]
        colnames(cleaned_data[[i]]) = colnames(cleaned_data[[1]])}
}
```

Binding all of the dataframes in the above list in one, and assigning column names

```
full_data = bind_rows(cleaned_data, .id = 'Year', )
colnames(full_data) = c('Year','Centre','Age','Conducted_Male','Passes_Male','Pass_Rate_Male',
                        'Conducted_Female','Passes_Female','Pass_Rate_Female',
                        'Conducted_Total','Passes_Total','Pass_Rate_Total')
```

Re-encoding missing values of centres

```
full_data$Centre[which(full_data$Centre == "",arr.ind = TRUE)] <- NA
full_data <- full_data %>% fill(Centre, .direction = 'down')

full_data[which(full_data == "..",arr.ind = TRUE)] <- NA
```

Re-encoding the years

```
for (i in 1:15) {
    full_data$Year[which(full_data$Year == i,arr.ind = TRUE)] <- 2006 + i}
```

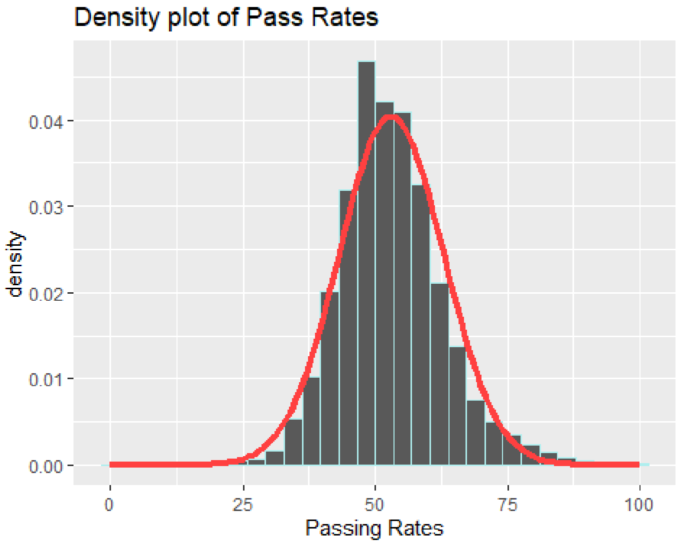
Year	Centre	Age	Conducted _Male	Passes_ Male	Pass_Rate_ Male	Conducted_ Female	Passes_ Female	Pass_Rate_ Female	Conducted _Total	Passes_ Total	Pass_Rate_ Total
2007	Aberdeen North										
2007	Aberdeen North	17	429	276	64.33566433	339	216	63.71681415	768	492	64.0625
2007	Aberdeen North	18	295	177	60	330	183	55.45454545	625	360	57.6
2007	Aberdeen North	19	147	87	59.18367346	177	102	57.62711864	328	190	57.92682926
2007	Aberdeen North	20	78	53	67.94871794	84	49	58.33333333	162	102	62.96296296
2007	Aberdeen North	21	69	38	55.07246376	85	49	57.64705882	154	87	56.49350649
2007	Aberdeen North	22	58	35	60.34482758	83	46	55.42168674	141	81	57.44680851
2007	Aberdeen North	23	52	28	53.8461538	62	31	50	115	59	51.30434782
2007	Aberdeen North	24	36	22	61.11111111	57	34	59.64912280	94	56	59.57446808
2007	Aberdeen North	25	58	36	62.06896551	70	39	55.71428571	128	75	58.59375
2007	Aberdeen North	Tota	1222	752	61.53846153	1287	749	58.19735819	2515	1502	59.72166998

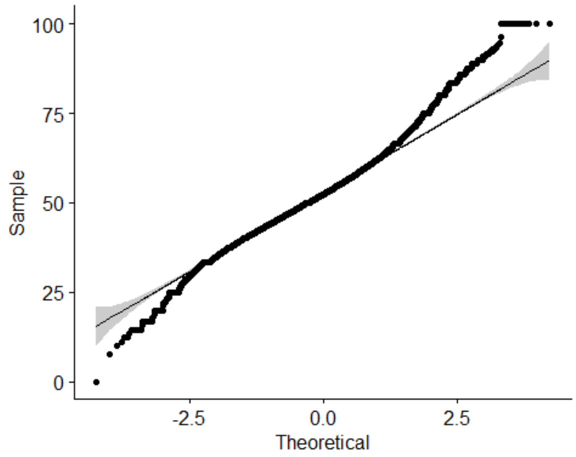
Exploring the data

```
Passrates <- as.numeric(full_data$Pass_Rate_Male)
mean(Passrates,na.rm = TRUE);var(Passrates,na.rm = TRUE)

## [1] 53.04242

## [1] 97.28735
```

Mean of the pass rates over the years is more than 50%, while the standard deviation seems high with a value of

```
ggplot(data.frame(Passrates), aes(x=Passrates)) +
geom_histogram(aes(y = after_stat(density)), col="paleturquoise",na.rm = TRUE) +
ggtitle("Density plot of Pass Rates") + xlab("Passing Rates") + stat_function(fun = dnorm, args = list(mean = mean(Passrates,na.rm = TRUE), sd = sd(Passrates,na.rm = TRUE)), lwd=1.5, col="brown1")
```





While the plots may look close to normal, doing some tests we see that the distribution is not normal.

In both the Andrerson-Darling and the Kolmogorov-Smirnov test, the null hypothesis is that the data follows a normal distribution.

```
library(nortest)
ad.test(Passrates)
##  Anderson-Darling normality test
##
## data:  Passrates
## A = 172.91, p-value < 2.2e-16

ks.test(Passrates,'pnorm')
## the Kolmogorov-Smirnov test
## Asymptotic one-sample Kolmogorov-Smirnov test
## data:  Passrates
## D = 0.99998, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

We can reject the null hypothesis for both the tests at 5% level of significance, and hence we cannot assume the Passing rates to be normally distributed.

To get the data for specific age and specific city, we create two functions:

Function for getting data of a specific city

```
specity <- function(city){
    citydata <- data.frame()

    x <- which(full_data== city,arr.ind = TRUE)
    p <- full_data[x[,1],]
    citydata <- rbind(citydata,p)
    rownames(citydata) <- NULL
    citydata <- citydata[citydata$Centre == city,]
    citydata <- citydata[!(citydata$Age == 'Total' | citydata$Age == ''),]
    return(citydata)
}
```

Function to get data for specific Age

```
specage <- function(df,age){
    agedata <- data.frame()
    x <- which(df$Age== age,arr.ind = TRUE)
    p <- df[x,]
    agedata <- rbind(agedata,p)
    rownames(agedata) <- NULL
    return(agedata)
}

nearLSE <- specity('Wood Green (London)')
LSEaged <- specage(nearLSE,25)
```

Year	Centre	Age	Conducted_Male	Passes_Male	Pass_Rate_Male	Conducted_Female	Passes_Female	Pass_Rate_Female	Conducted_Total	Passes_Total	Pass_Rate_Total
2007	Wood Green (London)	25	186	88	47.3118279	163	72	44.1717791	349	160	45.8452722
2008	Wood Green (London)	25	51	31	60.7843137	45	16	35.5555555	96	47	48.9583333
2009	Wood Green (London)	25	183	81	44.2622950	229	82	35.8078602	412	163	39.5631067
2010	Wood Green (London)	25	168	76	45.2380952	171	82	47.9532163	339	158	46.6076696
2011	Wood Green (London)	25	102	34	33.3333333	137	47	34.3065693	239	81	33.8912133
2012	Wood Green (London)	25	151	81	53.6423841	185	72	38.9189189	336	153	45.5357142
2013	Wood Green (London)	25	173	77	44.5086705	245	77	31.4285714	418	154	36.8421052
2014	Wood Green (London)	25	204	83	40.6862745	269	84	31.2267657	473	167	35.3065539
2015	Wood Green (London)	25	173	70	40.4624277	218	67	30.7339449	391	137	35.0383631

We see that data is missing for years past 9, as city name is different being Wood Green. So, we will combine this data too.

```
extra <- specity('Wood Green')
nearLSE <- rbind(nearLSE,extra)
```

Similarly, we can get the data for Nottingham (Colwick)

```
nearhome <- specity('Nottingham (Colwick)')
homeaged <- specage(nearhome,25)
```

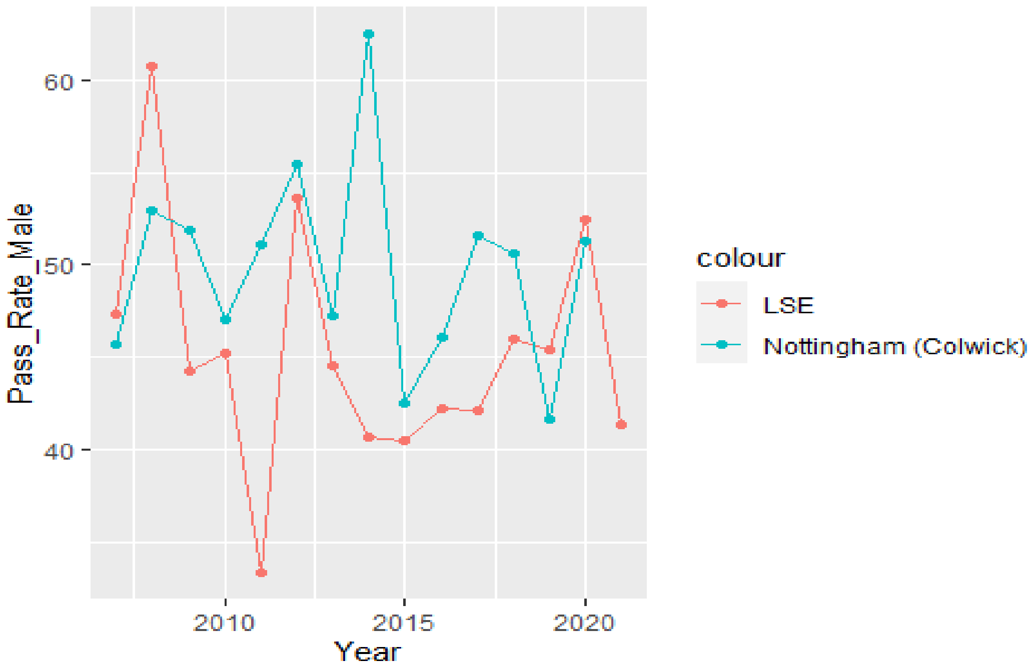
Difference between the two cities

Converting the data to numerical type,

```
lseagednum <-data.frame(sapply(LSEaged,function(x) as.numeric(as.character(x))))

homeagednum<-data.frame(sapply(homeaged,function(x) as.numeric(as.character(x)))
)
```

Plotting the Pass Rates

```
ggplot(data = lseagednum,aes(Year,Pass_Rate_Male, col = 'LSE')) +  geom_point()
+ geom_line() + geom_point(data = homeagednum, aes(Year,Pass_Rate_Male,col = 'No
ttingham (Colwick)')) + geom_line(data = homeagednum,aes(Year,Pass_Rate_Male,col
= 'Nottingham (Colwick)'))
```



Even though we notice no real trend in the passing rates of the two centres over the years, for most of the years, the Nottingham centre has had a higher passing rate than the Wood Green centre.

Permutation test

A permutation test nonparametric method for testing if two distributions are the same. It is particularly appealing when sample sizes are small, as it does not rely on any asymptotic theory. We will use this test to check whether the data of Males aged 25 in Nottingham (Colwick) and Wood Green (London) come from the same distribution, and thus will give us evidence if they differ significantly.

$$Ho : F_x = F_y \text{ versus } H_1 : F_x = F_y.$$

```
set.seed(1111)
x <- as.numeric(homeaged$Pass_Rate_Male)
y <- as.numeric(LSEaged$Pass_Rate_Male)
z <- c(x,y)
stat <- abs(mean(x) - mean(y))
k <- 0
for (i in 1:10000) {
    zperm <- sample(z,29)
    statperm <- abs(mean(zperm[1:14])-mean(zperm[15:29]))
    if (statperm > stat) k <- k+1}

pval <- k/10000
pval

## [1] 0.0479
```

Since p-value is less than 0.05, we can reject the null hypothesis that the samples are from the same distribution.

Now, we definitely know that there is a difference in the Passing rate of both the cities.

Logistic Regression Model.

To create our dataset, we created a **duplicate** function that repeated the a row the number of times people passed (with value 1), and failed(with value 0). Then we used that function to **Map** on our original dataset and hence, combined all of the data.

Our final dataset looks something like this:

To adjust for the high value of years affecting the model, we will encode the Years as 1,2,.....15

```
logisticdat$Year <- as.numeric(factor(logisticdat$Year))
model <- glm(pass_fail~.,data = logisticdat)
summary(model)
## Call:
## glm(formula = pass_fail ~ ., data = logisticdat)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.5684  -0.4933  -0.4112   0.4947   0.5971
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.6090856  0.0174409  34.923  < 2e-16 ***
## Year        -0.0041703  0.0004944  -8.435  < 2e-16 ***
## Centre       0.0611881  0.0041574  14.718  < 2e-16 ***
## Age         -0.0057452  0.0008091  -7.101 1.25e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.2483653)
##
##     Null deviance: 15013  on 60060  degrees of freedom
## Residual deviance: 14916  on 60057  degrees of freedom
## AIC: 86795
##
## Number of Fisher Scoring iterations: 2
```

We can compute a chi-square statistic to test if a model is useful.

$\chi 2$ statistic : 15013 - 14916 = 97 for 60060 - 60057 = 3 degrees of freedom.

Since the statistic has a p-value less than 0.0001, we can reject the null hypothesis that the model is not useful.

Predicting the pass rate for our student,

```
test <- function(age,year,centre){
    Age <- c(age);Year <- c(year);Centre <- c(centre)
    test <- data.frame(Year,Age,Centre)
    return(test)}

predict.glm(model,test(25,16,1),se.fit = TRUE)

## $fit
##         1
## 0.4599189
##
## $se.fit
## [1] 0.006583998
##
```

```
## $residual.scale
## [1] 0.4983626
##

predict.glm(model,test(25,16,0),se.fit = TRUE)
## $fit
##         1
## 0.3987308
##
## $se.fit
## [1] 0.006166066
##
## $residual.scale
## [1] 0.4983626
```

Looking at the predicted values, we see a value of 1(passing) 0.4599 for Centre 1, i.e. the centre near home, Nottingham (Colwick) while the value of 1 is 0.3987 at the Wood Green (London) Centre.

CONCLUSIONS

We have concluded that XYZ should take the test at the centre of Nottingham (Colwick), as it has a higher estimated passing rate according to the logistic regression model. By estimating the passing rate by the Years, Age and the location, we have a model useful for prediction, according to the deviance analysis. While not taking the Gender was a choice, it could also have been a limitation for this approach. On the other hand, the binary response with an optimal threshold is not a convincing predictor in this model. Because in different combinations of age and gender, most of the passing rates are near 50%, which means that no matter what the threshold has been set, we would have had the error rate of prediction closed to 50%. Another limitation would be lack of testing of the model, with a test set to estimate its accuracy even if all our other tests, pointed in favour of Nottingham.