ST447 – Data Analysis and Statistical Methods:

Individual Project – 35093

Introduction:

```
ID =          # LSE Student ID
source("XYZprofile.r")
XYZprofile(ID)

## The profile of XYZ:
## - Age:  19
## - Gender:  Male
## - Home address:  Mill Hill (London)
```

XYZ, a **19**-year-old **Male** individual, perceives his driving skill to be average after learning driving for some time. He wants to take the practical car test in UK and is now contemplating between two options which is: either taking the test at the nearest test centre to his home which is at **Mill Hill (London)**, or to take it at the centre nearest to LSE which is Wood Green (London). The data analysis aims to provide XYZ some insight into the expected passing rates at these respective locations, and provide a recommendation for his strategic choice of test centre.

1.1 Data Source:

The dataset we use, 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 – 2023) and test centre.

In our data analysis using R, we consolidated information from multiple years (2007 – 2023) for the Mill Hill and Wood Green locations. The compiled dataset incorporates all age groups and genders from both locations. This comprehensive approach enables us to assess whether XYZ's specific characteristics, such as being a **19**-year-old **Male**, play a significant role in our decision regarding the choice of the test centre.

```
comp_data <- list() # creates an empty list to store data
list_of_sheets <- list_ods_sheets("dvsa1203.ods") # reads the names of sheets in the
ODS File
num <- length(list_of_sheets) # stores the no. of sheets in the file
for (x in 2:num){ # for loop to run through the sheets, reads the data in the ODS file
and stores the data required in an empty list
    data <- read_ods("dvsa1203.ods", sheet = list_of_sheets[x], skip = 5)
    colnames(data) <- c("CityName")
    demo <- which(!is.na(data$CityName), arr.ind = TRUE)
    for (i in 1:(length(demo)-1)){
    s_ind <- demo[i];   f_ind <- demo[i+1]-1
    c_name <- data$CityName[s_ind]
      for(j in s_ind:f_ind){
      data$CityName[j] <- c_name
    }}
    data <- data[, colSums(is.na(data)) < nrow(data)]
    colnames(data) <- c("CityName", "Age", rep(c("Male tests", "Female tests", "Total
tests"), each = 3))
    data$Year <- substr(list_of_sheets[x], 1, 4)
    comp_data[[list_of_sheets[x]]] <- data[data$CityName %in% c("Mill Hill (London)",
"Wood Green (London)", "Mill Hill", "Wood Green") , ]
    comp_data[[list_of_sheets[x]]] <- data[ , ]
}
```

```r
# Combining Data Frames
working_data <- data.frame()
for (i in 1:length(comp_data)){
  working_data <- rbind(working_data, comp_data[[i]])
}
working_data <- na.omit(working_data)
write.xlsx(working_data, "output.xlsx")
```

The given R code is used to compile all the data as mentioned earlier and is stored in another Excel Workbook named "output.xlsx." The data is further edited in the Excel sheet and is arranged as shown below.

```r
# Loading the data
setwd("C:/Users/General/Desktop/MSc Data Science/ST 447/Project Submission")
data_v2 = read.xlsx(xlsxFile = "output.xlsx", sheet = "working_data")
data_v2$Failed <- data_v2$Conducted - data_v2$Passed # calculates no. of Failed
colSums(is.na(data_v2)) # to find NA values in each column
```

```
##  CityName       Age    Gender Conducted    Passed Pass_Rate      Year    Failed
##         0         0         0         0         0         0         0         0
```

```r
head(data_v2)
```

```
##    CityName Age Gender Conducted Passed Pass_Rate Year Failed
## 1 Mill Hill  17   Male       592    257  43.41216 2007    335
## 2 Mill Hill  18   Male       480    172  35.83333 2007    308
## 3 Mill Hill  19   Male       357    144  40.33613 2007    213
## 4 Mill Hill  20   Male       286    111  38.81119 2007    175
## 5 Mill Hill  21   Male       249     88  35.34137 2007    161
## 6 Mill Hill  22   Male       238     86  36.13445 2007    152
```

1.2 Methods:

1.2.1 *Logistic Regression*

A. Data Collection

To ensure I had a comprehensive dataset, I considered the total number of people who passed the driving test and those who did not. Using this information, I expanded the dataset by creating multiple records for each row, corresponding to the total number of passed and failed. This approach allows us to capture a more detailed picture of individual experiences and provides a solid foundation for our analysis.

```r
# Expanding the Dataset

new_data_1 <- data_v2[rep(seq_len(nrow(data_v2)), times = data_v2$Passed),] # replicat
es each row as many times as the No. of Passed and stores "Pass" as the Result
new_data_1$Result <- "Pass"
new_data_2 <- data_v2[rep(seq_len(nrow(data_v2)), times = data_v2$Failed),] # replicat
es each row as many times as the No. of Failed and stores "Fail" as the Result
new_data_2$Result <- "Fail"
new_data <- rbind(new_data_1, new_data_2) # combines the data from the 2 data frames
new_data_check <- subset(new_data, select = c("CityName", "Age", "Gender", "Year", "Re
sult")) # subsets the data with only the columns we require
new_data_check$Result <- ifelse(new_data_check$Result == "Pass", 1, 0) # changes the
target variable to 1 if its Pass or 0 if its Fail
str(new_data_check) # shows the data type of the columns in the dataframe
```

```
## 'data.frame':    132697 obs. of  5 variables:
##  $ CityName: chr  "Mill Hill" "Mill Hill" "Mill Hill" "Mill Hill" ...
##  $ Age     : num  17 17 17 17 17 17 17 17 17 17 ...
##  $ Gender  : chr  "Male" "Male" "Male" "Male" ...
```

```
##  $ Year    : num  2007 2007 2007 2007 2007 ...
##  $ Result  : num  1 1 1 1 1 1 1 1 1 1 ...
```

B. **Assumptions**

As part of the exploratory analysis, I generated graphs to visualize trends in the data. Based on these visualizations, there is an initial indication that Age and Year exhibit a linear relationship with the Pass Rate. These assumptions are derived from the observed patterns in the graphs, suggesting a linear relation between these factors and the likelihood of passing the driving test. Thus, in the Logistic Regression, I do not consider Age and Year as a factor and use it as a linear variable.

```
# Checking Yearly Trend
yearly_data <- subset(data_v2, Age == 19 & Gender == "Male", select = c("CityName", "Age", "Gender", "Passed", "Failed", "Year", "Pass_Rate"))
ggplot(data = yearly_data, aes(Year, Pass_Rate, color = CityName)) + geom_point() + geom_line() + geom_smooth(method = "glm", se = FALSE) # Plot in Fig 1

# Checking Age Data
age_data <- subset(data_v2, Year == 2022 & Gender == "Male", select = c("CityName", "Age", "Gender", "Passed", "Failed", "Year", "Pass_Rate"))

ggplot(data = age_data, aes(Age, Pass_Rate, color = CityName)) + geom_point() + geom_line() + geom_smooth(method = "glm", se = FALSE) # Plot in Fig 2
```

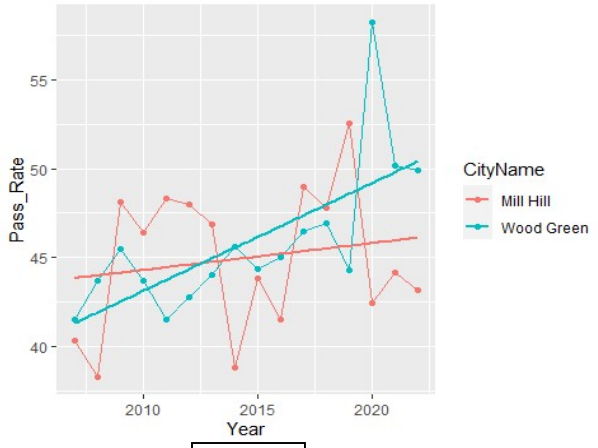


Fig 1

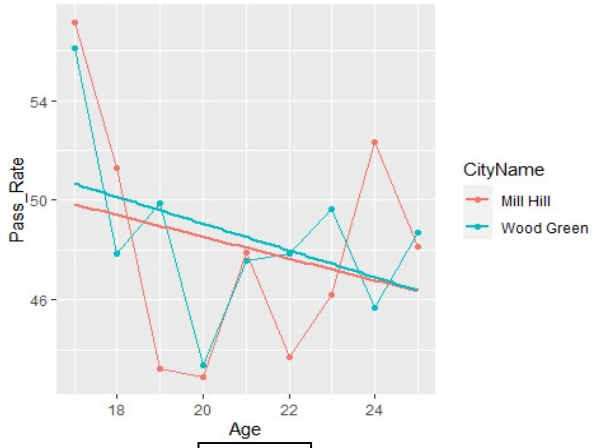


Fig 2

C. **Procedure**

I opted for multinomial logistic regression to dig into our data, since we are keen on predicting whether someone passes or fails the driving test. I take the Result which is a binary variable – **Pass (1) or Fail (0)** as our target variable (Y), the variable we want to predict. The dependent variables (X) are the factors that can influence our prediction and I take **City** where they are giving the test, **Age** and **Gender** of XYZ and the **Year** in which they are taking the test as the dependent variables. This helps us figure out how these factors link to the chances of passing or failing the driving test.

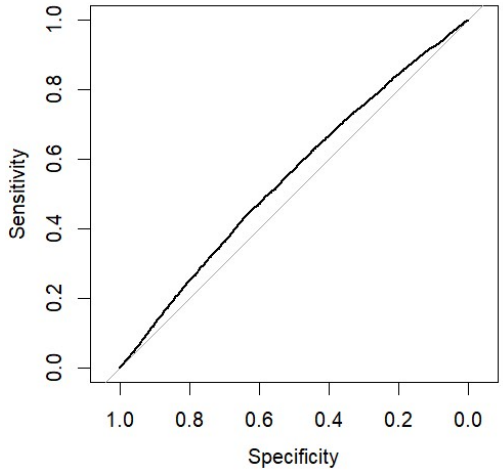
```
set.seed(12) # We are splitting our data into train and test to check the accuracy of our model later (we randomly sample 80% of our data as training to train the model)
training.data <- sample(nrow(new_data_check), size = 0.8*nrow(new_data_check))
training_set <- new_data_check[training.data, ]
```

```
# Multinomial Logistic Regression
glm_lr1_1 <- glm(Result ~ . , data = new_data_check, family = 'binomial', subset = tra
ining.data)

# We take the equation as Result ~ CityName + Age + Gender + Year
```

D. **Results**

```
summary(glm_lr1_1)$coef # gives the summary of the logistic regression
## Call:
## glm(formula = Result ~ ., family = "binomial", data = new_data_check,
##     subset = training.data)
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -36.541015   2.678374 -13.643  < 2e-16 ***
## CityNameWood Green  -0.024855   0.012539  -1.982   0.0474 *
## Age                 -0.015376   0.002489  -6.177 6.52e-10 ***
## GenderMale           0.246366   0.012505  19.702  < 2e-16 ***
## Year                 0.018100   0.001330  13.612  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1



# Testing Set (to test the accuracy of the data)
testing_set <- new_data_check[-training.data,]
testing_set$res <- predict(glm_lr1_1, testing_set, type = "response") #uses the model
to predict the probability of Passing based on the dependent variables in the test set
testing_set$predicted.classes <- ifelse(testing_set$res > 0.5, 1, 0) # if-else stateme
nt to check the prediction, 1 if the prob > 0.5 else 0
mean(testing_set$predicted.classes == testing_set$Result) # to check Model Accuracy

## [1] 0.5773173

ts_fac <- (testing_set) # To create Confusion Matrix we create a new dataframe
ts_fac$predicted.classes <- factor(ts_fac$predicted.classes)
ts_fac$Result <- factor(ts_fac$Result)
confusionMatrix(ts_fac$Result, ts_fac$predicted.classes)

## Confusion Matrix and Statistics
##           Reference
## Prediction     0     1
##          0 15032   308
##          1 10910   290
##                Accuracy : 0.5773
##                  95% CI : (0.5713, 0.5833)


roc_curve <- roc(ts_fac$Result, ts_fac$res)
plot(roc_curve) # Fig 3
auc(roc_curve)
## Area under the curve: 0.5484
```

Fig 3



E. **Interpretation**

From the summary function we get the coefficient estimates, p-value and significance levels for each factor, as to how they are likely to affect the Result being Pass (1) or Fail (0).

- The value of Estimate are the beta coefficient estimates associated with each dependent variable.

- The Standard Error represents the accuracy of the Estimates, the smaller it is the more confident we are about the estimate
- The Z-Statistic is the coefficient estimate divided by the standard error and Pr(>|z|) is the p-value corresponding to the Z-Statistic. The smaller the p-value, the more significant our estimate is.

In our model, the estimate for CityName shows that the estimated log-odds of the result decreases by 0.0248 if the person is from Wood Green, and the * indicates that this change is somewhat significant (p-value < 0.05). Similarly for Age, as it increases there is a decrease in the estimated log-odds of the result by 0.0154 and this change is statistically significant since p-value is less than 0.001. For Gender and Year, the estimated log-odds of the result increase by 0.2464 if the person is a Male and by 0.0181 with the increase in Year. They are also highly significant as indicated by the *** and p-value < 0.001. Thus, these coefficients show how each factor affects the likelihood of our result and the p-value tells us if these effects are significant and due to the real differences in our data or if they are random. The lower the p-value, the more confident we are in the Result.

The Confusion Matrix shows us how well our predictions match the actual outcomes. 290 correct positive predictions (True Positives) and 15032 correct negative predictions (True Negatives). However, there were 308 instances of incorrectly predicting positive outcomes (False Positives) and 10910 instances of incorrectly predicting negative outcomes (False Negatives) and thus the overall accuracy was only 57.73% indicating the proportion of correct predictions. The 95% Confidence Interval suggests that we are confident that the true accuracy falls between the range of 57.13% and 58.33%.

The ROC (Receiver Operator Characteristic) is another way to determine the accuracy of our model at a threshold value defined by us. The model's accuracy is determined by the Area Under the Curve (AUC) and higher the area, the better the model. Our aim is to push the curve towards the left corner and higher the curve, better the model. Our AUC score is 0.5484, suggesting room for improvement and our current curve also indicates opportunities to enhance our model.

F. Analysis

Our analysis delved into various factors which would influence the Result of the driving test and revealed insights into the passing rates at the two test centre choices. The factors which I used for our logistic regression model had statistically significant coefficient estimates and we now address the three pivotal questions posed by XYZ, the expected passing rates at the two test centres and the most strategic location for a successful driving test experience.

```
check <- data.frame(CityName = c("Mill Hill", "Wood Green"), Age = 19, Gender = "Male"
, Year = 2023) # We create a dataframe with the 2 test centres and XYZ's profile
check$Result <- predict(glm_lr1_1, newdata = check, type = "response")
check  # We predict the probability of XYZ Passing at each of the test centre
```

```
##     CityName Age Gender Year    Result
## 1  Mill Hill  19   Male 2023 0.5072955
## 2 Wood Green  19   Male 2023 0.5010823
```

This gives us the expected passing rates of XYZ at nearest test centre to his home and nearest test centre to LSE.

The **Expected Passing Rate at Mill Hill (nearest test centre to XYZ's home) is 50.729**.

The **Expected Passing Rate at Wood Green (nearest test centre to LSE) is 50.108**.

In conclusion to this, **I would suggest XYZ to take the test at the centre near Mill Hill, which is the nearest test centre to XYZ's home**. Although there isn't much difference in the two expected passing rates, given the expected passing rate at Mill Hill is higher, XYZ would have a better chance to have a successful outcome at his driving test endeavour.

G. **Limitations**

While logistic regression offers valuable insights into predicting Pass or Fail outcomes, acknowledging its imperfections is crucial. Our model, while informative, falls short of an ideal accuracy score, revealing the inherent challenges in capturing the complexities of real-world driving test predictions.

Also, our less-than-optimal ROC curve and AUC score underscores the struggle in distinguishing between passing and failing with high precision. These challenges underscore the need for a nuanced understanding, emphasizing that while logistic regression provides good insights, sometimes it may not fully grasp the intricacies of the relationships between the different variables.

In conducting our analysis, I focused on specific age and gender groups, to assume a linear relationship between years and the likelihood of passing the test across both locations. This simplification, while aiding our analysis, may impact the accuracy of predicting test outcomes. Additionally, the reliance on data from a single year and a particular gender to assume the linear relationship for age may also introduce some limitations. Recognizing these constraints, it is essential to acknowledge that our model might not fully capture the diverse dynamics at play, and further refinements could enhance its predictive capabilities.

Exploring non-linear relationships between age, year, and the likelihood of passing the test and considering different interaction terms in the model may help us improve the limitations and capture more nuanced patterns.

More advanced modelling like decision trees, random forests or gradient boosting may help us capture complex relationships between the various factors and improve the prediction accuracy.

1.2.2 *Wald Test for Equality of Means*

A. **Data Collection**

```
data_MH <- subset(data_v2, CityName == "Mill Hill") # data for Mill Hill
data_WG <- subset(data_v2, CityName == "Wood Green") # data for Wood Green
```

I subset the data from the original dataset, on the basis of the City names and store it in two different data frames.

B. **Procedure**

I am using the Wald test to examine if there is a statistically significant difference in the passing rates between two locations. This test directly assesses the passing rates of the data from both places, aiming to check significant disparities in the passing rates. In essence, the Wald test helps us determine if one location exhibits a significantly better or worse passing rate compared to the other, allowing for a data-driven suggestion regarding the comparative evaluation of the two locations. I choose the mean of the passing rates of both the locations as the estimator for ease of interpretation and it serves as a metric to compare the passing rates of the two locations.

We consider,

Null Hypothesis (H_0):

H_0 : The mean passing rates at Mill Hill and Wood Green are equal $\rightarrow \mu_x - \mu_y = 0$

vs

Alternate Hypothesis (H_1):

H_1 : The mean passing rates at Mill Hill and Wood Green are significantly different (are not equal) $\rightarrow \mu_x - \mu_y \neq 0$

$\mu_x \rightarrow$ Mean Passing Rate at Mill Hill; $\mu_y \rightarrow$ Mean Passing Rate at Wood Green

```
mu1 = mean(data_MH$Pass_Rate); mu2 = mean(data_WG$Pass_Rate) # Mean of Pass Rate
Sx = var(data_MH$Pass_Rate); Sy = var(data_WG$Pass_Rate) # Sample Variance - Pass Rate
```

```
n1 = length(data_MH$Pass_Rate);n2 = length(data_WG$Pass_Rate) #no. of obs of Pass Rate
var = (Sx/n1) + (Sy/n2) # Pooled Variance since both variances are not equal
se = sqrt(var) # Standard Error which is the Estimator for Standard Deviation
T_stat <- (mu1 - mu2) - 0 / se # T-statistic follows Standard Normal Distribution
```

C. **Results**

```
T_stat
## [1] 0.06106689
```

We calculate the T- Statistic as 0.061 for our Wald test for Equality of Means. Since it is a two – tailed test, to check whether there is any significant difference in the two means, the value of Z-Statistic at 95% Confidence Interval is determined by $Z_{0.05} = 1.96$.

D. **Analysis**

We shall Reject H_0 if the absolute value of our T-Statistic is greater than the Z-Statistic value.

Since $0.061 < 1.96$, we **DO NOT REJECT H_0**.

Thus, there is **no significant difference** in the **means of the passing rates at Mill Hill or Wood Green**.

Based on this method, XYZ can take the test at either of the driving centres since there is no significant difference in the passing rates at Mill Hill or Wood Green.

E. **Limitations**

We may encounter some limitations when we apply the Wald Test to small sample sizes, as it relies on asymptotic normality assumptions, potentially leading to inaccurate results. Also, some factors influencing the passing rates may not be adequately captured, compromising the reliability of the test.

1.3 Conclusion

In conclusion, the method of logistic regression revealed some differences in passing rates between two locations, with a marginal advantage suggested for one over the other. However, upon subjecting the data to the Wald test, the observed variance in mean passing rates did not reach statistical significance. This shows the need for a better interpretation – while logistic regression provides directional insights, the Wald test emphasizes the importance of statistical rigor. As we navigate this analytical landscape, it is crucial to consider both the practical and statistical implications, recognizing the potential influence of unexplored factors.

I would recommend XYZ to give his driving test at the test centre nearest to his house in Mill Hill since the logistic regression shows a better expected passing rate for Mill Hill. However, considering XYZ's average driving skill and the lack of statistical significance in the difference in passing rates for both Mill Hill and Wood, XYZ could also give his driving test at the test centre nearest to LSE at Wood Green.