Individual Project

ST447 Data Analysis and Statistical Methods

37521

4 December 2020

```r
# libraries used
library(tidyverse)
library(readODS)
source(’XYZprofile.R’)
```

```
## The profile of XYZ:
## - Age:  22
## - Gender:  Male
## - Home address:  Luton
```

Introduction

XYZ can take his driving test either at the test center in Wood Green (nearest to LSE) or Luton (nearest to home). The goal of this project is to give XYZ a recommendation based on historical data.

My analysis consists of two steps. First, I use logistic regression to calculate my friend’s expected passing rate for both Wood Green and Luton. I then use a Wald test to test whether there is indeed a statistically significant difference in the passing rates for 22 year-old, male test takers.

```r
sheets = list_ods_sheets(’data/dvsa1203.ods’)[-1]

extractData = function(sheet, city){
  # function for extracting the data from the ODS file
  df = read_ods(’data/dvsa1203.ods’, sheet, skip = 6)
  index = which(df[,1]==city)
  data = df[(index+1):(index+9),-c(1)]
  data = map_dfc(data, as.numeric)
  men = data[,1:4]
  men$Gender = ’m’
  women = data[,c(1,5:7)]
  women$gender = ’f’
  names(women) = names(men)
  data = rbind(men, women)
  data$Year = sheet
  data$City = city
  return(data)
}
```

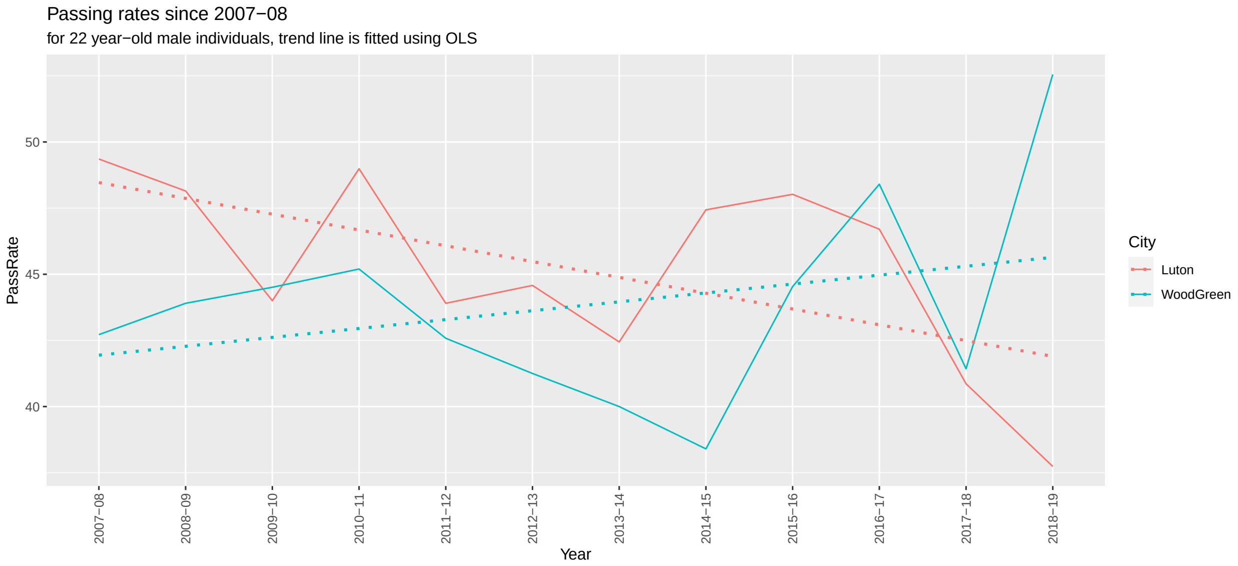
```
Luton = map_dfr(sheets, extractData, ’Luton’)
WoodGreen1 = map_dfr(sheets[1:6], extractData, ’Wood Green (London)’)
WoodGreen2 = map_dfr(sheets[7:12], extractData, ’Wood Green’)
WoodGreen = rbind(WoodGreen1, WoodGreen2)
WoodGreen$City = ’WoodGreen’
data = rbind(Luton, WoodGreen)
names(data) = c("Age", "Conducted", "Passes", "PassRate", "Gender", "Year", "City")
saveRDS(data, file = "data/LutonWooGreenData.rds")
```

Data Wrangling

For my analysis, I only use data from the previous three years. This is a subjective choice that is based on a balance between a large enough sample size and using only the most recent data. The reason behind the latter is that passing rates can change over time. In fact, traffic conditions often gradually change in an area, so do driving instructors and cars at a test center, and possibly many other factors. In other words, driving tests at a certain site can become more or less challenging over time. This claim is also supported by the following plot. Generally speaking, we can see that the passing rates in Wood Green and Luton have developed in opposite directions.

```
sheets = list_ods_sheets(’data/dvsa1203.ods’)[-1]

# plot for the passing rates
data %>%
  filter(Age==22, Gender==’m’) %>%
  group_by(Year, City) %>%
  summarise(PassRate = PassRate) %>%
  ggplot(aes(Year, PassRate, group=City, color=City))+
    geom_line() +
    geom_smooth(method = ’lm’, se = FALSE, linetype=’dotted’) +
    theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1)) +
    labs(title = ’Passing rates since 2007-08’,
         subtitle = ’for 22 year-old male individuals, trend line is fitted using OLS’)
```



To fit the logistic regression model later, the data must first be reformatted. More specifically, it should be a matrix in which each row represents one individual test taker and the columns represent the features we have at hand, i.e. **Age** (ranges from 17 to 25), **Gender** (binary, either female or male), and **City** (also binary, either Wood Green or Luton). In addition, there is one column for the response variable which indicates whether the individual has passed their driving test or not (binary, 1 for passed and 0 for not passed).

```
# reformatting of the data
counts = data %>%
  filter(Year %in% sheets[1:3]) %>%
  group_by(Age, Gender, City) %>%
  summarise("TotalPasses" = sum(Passes),
            "TotalFails" = sum(Conducted-Passes))

IndPass = uncount(counts, TotalPasses)[-4]
IndPass$Pass = 1
IndFail = uncount(counts, TotalFails)[-4]
IndFail$Pass = 0

PassData = rbind(IndPass, IndFail)
PassData %>%
  group_by(City) %>%
  summarise(n=n())
```

```
## # A tibble: 2 x 2
##   City          n
##   <chr>     <int>
## 1 Luton     17991
## 2 WoodGreen 10956
```

There are $n = 10,956$ and $m = 17,991$ individuals who took their test in Wood Green and Luton, respectively.

Logistic Regression

Regression is an natural choice for finding the expected passing rate for my friend. Note that the regression function is defined as the conditional expectation $r(X) = E(Y|X)$. In the case of a Bernoulli response variable, $E(Y = 1|X) = P(Y = 1|X)$, i.e. the conditional expectation is also the probability of $Y = 1$ given X.

Consequently, we could use plain linear regression to derive the expected passing rate using Age, Gender, and City as predictors (this is equivalent to a classification task). However, linear regression is not well suited to model a binary response variable. For example, it can predict probabilities that are negative or larger than 1. A more sensible approach is to use logistic regression which uses the logit link function to map a linear combination of features into the $(0, 1)$ range. This can then be interpreted as a probability, or in our case as the expected probability of a *Pass* given an individual's features.

Using logistic regression, we model the probability of passing the driving test given the three features as follows.

$$P(Pass = 1|Gender, Age, City) = \frac{\exp(\beta_0 + \beta_1 Age + \beta_2 Gender + \beta_3 City)}{1 + \exp(\beta_0 + \beta_1 Age + \beta_2 Gender + \beta_3 City)}$$

Here, both Gender and City are dummy variables, i.e. $Gender = 1$ stands for male and 0 for female, and $City = 1$ stands for Wood Green and 0 for Luton. Age, by contrast, is a continuous variable. We rely on R to fit the coefficients $(\beta_0, \beta_1, \beta_2, \beta_3)$.

```
logreg.fit = glm(Pass~Age+Gender+City, data = PassData, family = ’binomial’)
summary(logreg.fit)
```

```
##
## Call:
## glm(formula = Pass ~ Age + Gender + City, family = "binomial",
##     data = PassData)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.1548  -1.0240  -0.9183   1.3170   1.4983
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -0.313345   0.104055  -3.011  0.00260 **
## Age           -0.016609   0.005071  -3.275  0.00106 **
## Genderm        0.306685   0.024214  12.666  < 2e-16 ***
## CityWoodGreen  0.235536   0.025030   9.410  < 2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 38956  on 28946  degrees of freedom
## Residual deviance: 38698  on 28943  degrees of freedom
## AIC: 38706
##
## Number of Fisher Scoring iterations: 4
```

We can see that all coefficients are significant. More specifically, we can see that Age has a negative and Gender a positive coefficient, indicating that both older and female test takers fail more often - this may be a sign of unfair discrimination against women and older test takers, although drawing a definite conclusion here would require a more in-depth analysis (also note that we can not see if one test center discriminates more than the other here).

Of course, we are particularly interested in the City variable since it is the one variable we can influence. Its coefficient is positive, indicating that taking the test at site Wood Green is associated with a higher probability of passing. In fact, we can say that taking the test in Wood Green increases the log-odds of passing by 0.236 as compared to Luton, keeping all other variables fixed. This interpretations results from the fact that the previous equation can be rewritten in terms of the log-odds, i.e.

$$log\frac{P(Y=1|Age, Gender, City)}{P(Y=0|Age, Gender, City)} = \beta_0 + \beta_1 Age + \beta_2 Gender + \beta_3 City$$

A 95 percent confidence interval for the City coefficient is (0.186, 0.285). Therefore, we can be confident that the effect is indeed greater than zero.

```
confint(logreg.fit)
```

```
##                     2.5 %      97.5 %
## (Intercept)   -0.51727612 -0.10937751
## Age           -0.02655446 -0.00667499
## Genderm        0.25924037  0.35415785
## CityWoodGreen  0.18647442  0.28459280
```

Finally, all that remains is to calculate **my friend’s expected passing rate** for both sites. We fix the values for Age and Gender but vary the City variable inside R’s predict function. We find that there is a delta of 5.79 percentage points in favor of Wood Green.

```
LutonEPR = predict(logreg.fit, data.frame(Age=22, Gender=’m’, City="Luton"),
                   type = "response")
WoodGreenEPR = predict(logreg.fit, data.frame(Age=22, Gender=’m’, City="WoodGreen"),
                       type = "response")

cat("XYZ’s expected Passing Rate in\n- Wood Green:", WoodGreenEPR, ’\n- Luton:’, LutonEPR)
```

```
## XYZ’s expected Passing Rate in
## - Wood Green: 0.4659203
## - Luton: 0.4080418
```

Wald test

In addition, I also test whether there is indeed a statistically significant difference in the passing rates for 22 year-old males. Accordingly, I restrict the test to just the relevant age group and gender.

Let $W_1...W_n \sim Ber(\theta_W)$ and $L_1...L_m \sim Ber(\theta_L)$ be Bernoulli random variables that correspond to test takers at sites Wood Green and Luton, respectively. $W, L \in \{1 \text{ if pass}, 0 \text{ if fail}\}$ are assumed to be independent from each other. We want to test the null hypothesis $H_0 : \theta_W - \theta_L = 0$ against the alternative hypothesis $H_1 : \theta_W - \theta_L \neq 0$. A natural estimator for the parameters is the sample average, i.e. $\hat{\theta}_W = \bar{W}$ and $\hat{\theta}_L = \bar{L}$, which are also their MLEs.

We can then use a Wald test to test equality of the means. Importantly, I use a two-sided test to ensure that the result is not distorted by my knowledge of the data so far. In addition, since the stakes are so low, we will be satisfied with a test at significance level 0.1. That is, the probability of incorrectly rejecting H_0 (type-II error) should be no more than 10 percent. We will therefore reject the null-hypothesis if the test statistic is larger then 1.64 (95 percentile of $N(0, 1)$), or equivalently, if the p-value is smaller than 0.10.

The test statistic T and the standard error are calculated as follows,

$$T = \frac{\hat{\theta}_W - \hat{\theta}_L}{SE(\hat{\theta}_W - \hat{\theta}_L)} \sim N(0, 1)$$

$$SE(\hat{\theta}_W - \hat{\theta}_L) = \sqrt{\frac{\hat{\theta}_W(1 - \hat{\theta}_W)}{n} + \frac{\hat{\theta}_L(1 - \hat{\theta}_L)}{m}}$$

where $n = 465$ and $m = 580$ are the total number of 22 year-old, male test takers in Wood Green and Luton, respectively.

The observed difference $\hat{\theta}_W - \hat{\theta}_L$ is 0.05975 and the calculated standard error is 0.0309. As a result, the test statistic T is 1.93, which corresponds to a p-value of 0.053. Consequently, we come to the conclusion that the test statistic is an *extreme or unlikely value* on the positive side of the bell curve, indicating that θ_W is indeed larger.

```
WoodGreen = filter(PassData, City==’WoodGreen’ & Age == 22 & Gender == ’m’)$Pass
Luton = filter(PassData, City==’Luton’ & Age == 22 & Gender == ’m’)$Pass

n = length(WoodGreen); m = length(Luton)
W = mean(WoodGreen); L = mean(Luton)
```

```
d_hat = W-L
SE = sqrt(W*(1-W)/n + L*(1-L)/m)
T_ = d_hat/SE
p_value = 2-2*pnorm(T_)

cat(’d_hat =’, d_hat, ’\nSE =’, SE, ’\nT =’, T_, ’\np-value =’, p_value)
```

```
## d_hat = 0.05975158
## SE = 0.03090385
## T = 1.933467
## p-value = 0.05317867
```

As a result, we reject the null hypothesis as there is significant evidence that the passing rates for 22-year old males differ between the two sites.

Conclusion and limitations

Both the logistic regression and the Wald test arrive at the same conclusion. **My suggestion is therefore: take the test in Wood Green!**

Finally, I want to discuss some of the limitations of this analysis. To begin with, I do not take into account how long ago each test was taken. Instead, equal weight was given to each individual. This issue is somewhat mitigated by the fact that I use only the three most recent years. Nevertheless, an improvement may be possible by weighting individuals based on the year. We may also consider to include the season or even weather conditions of the test day as predictors, if we can obtain such information.

Secondly, I want to point out that the logistic regression and the Wald test are based on slightly different underlying assumptions as well as a different subset of the data. Logistic regression implicitly assumes that all $Y_i|Age_i, Gender_i, City_i \sim Bernoulli(\theta_i)$. That is, we model all test takers with a single model. For the Wald test, by contrast, I only take into account all 22 year-old male individuals and assume that $W_i|(Age = 22, Gender = m) \sim Bernoulli(\theta_W)$ and $L_j|(Age = 22, Gender = m) \sim Bernoulli(\theta_L)$. However, the two methods are complimentary and both were crucial to my final recommendation - while the logistic regression provides two concrete expected passing rates, the Wald test tells us whether the difference is statistically significant. In fact, that the two methods arrived at the same conclusion indicates that the recommendation is robust.

In addition, some would perhaps criticize that the significance level used for the Wald test ($\alpha = 0.1$) is too high. Indeed, the most used values in practice are 0.05 and 0.01. However, as mentioned previously, there is no reason to be any more restrictive than necessary here in my opinion as the consequences of falsely rejecting the null-hypothesis are harmless.

Lastly, age is represented as a continuous variable even though the oldest age group (25 year-olds) comprises all individuals aged 25 or beyond. That is, it presumably includes a small group of much older individuals with perhaps very different passing rates. This may inflate the City coefficient in the logistic regression model.