Testing Through Data

ST447 FINAL PROJECT

86466 | Data Analysis & Statistical Methods | 31st January 2018

Table of Contents

Introduction

XYZ, a friend of mine, has been learning to drive for some time and now feels ready to take the driving test. There are two sensible locations for her to take the test at: the one nearest to her home, or the one nearest to her university (LSE) which is at Wood Green (London). Knowing that I am a Statistics student (also studying at LSE), XYZ has asked me to help her decide which test centre would give her the best chance of passing.

This report aims to arrive at a decision using simple data analysis, data visualisation & a statistical proof, validating my decision (via the chosen method of ***linear regression)***.

In the midst of reaching my conclusion, I will also provide the steps used for ***data pre-processing*** (which basically means transforming raw, messy data into an understandable format) as well as the R code used at each stage of my calculation so it can be easily replicated.

Before we proceed, the profile for my friend has been generated below:

```
## The profile of XYZ:
## - Age:  22
## - Gender:  Female
## - Home address:  Bury (Manchester)
```

For clarity & concreteness, I am assuming that my friend's name is Sarah Parker.

Data Collection & Pre-Processing

The data used for this analysis **DVSA1203** is available at https://www.gov.uk/government/statistical-data-sets/car-driving-test-data-by-test-centre. It contains information on car pass rates by age (17 to 25-year olds), gender, year (2007-2014) and test centre.

I have converted the file into a .csv format and read it into excel; renaming of the Columns and addition of another variable called "Year" have also been done. Since the pre-processing code is quite extensive, I will demonstrate the code used for this step only for the year 2014 from Bury (Manchester); it can be generalised to the other years and for Wood Green (London) as well. (The entire code for each pre-processing step can be viewed at http://rpubs.com/Akeel23/353912. This file also contains my ***complete*** code used for this project with explanations)

1)Read in Data from year 2007-2014 as CSV:

```
sheet_1_201415 <- read.csv ("sheet_1.csv", stringsAsFactors = FALSE,header = T )
```

2) Rename columns for each sheet from each year:

```
colnames(sheet_1_201415) <- c("Test_Centre","Age","Male_Tests_Conducted", "Male_Passes", "Male_Pass_Rate","Female_Tests_Conducted", "Female_Passes",
```

```
"Female_Pass_Rate", "Total_Tests_Conducted", "Total_Passes", "Pass_Rate")
```

3) Extract Data for Tests conducted in Bury (Manchester) and Wood Green (London) from years 2007 -2014:

```
bury.2014 = sheet_1_201415[sheet_1_201415$`Test_Centre`=="Bury
(Manchester)" , ]
```

4) I then merge all these data frames together to create one single data frame for each Area:

```
df_bury<- rbind(bury.2007, bury.2008, bury.2009, bury.2010, bolton.2
011, bury.2012, bury.2013, bury.2014)
df_wgreen<- rbind(wgreen.2007, wgreen.2008, wgreen.2009, wgreen.2010
, wgreen.2011, wgreen.2012, wgreen.2013, wgreen.2014)
```

Assumptions

- Since there is no data for Bury (Manchester) in 2011-12, I am taking the data for Bolton, which is the next closest Test Centre to Bury of approximately 6 miles (according to www.dft.gov.uk/fyn/practical.php).
- Sarah cannot wait until she is at least a year older in order to take the test (explanation for this assumption provided later).
- Places marked “Bury” and “Bury (Manchester)” are the same. Note that the places named “Bury” have been renamed to the former for ease of entry of formulae in R.
- Places marked “Wood Green” and “Wood Green (London)” are the same. Note that the places named “Wood Green” have been renamed to the former for ease of entry of formulae in R.
- The basic assumptions for linear regression to be valid, which are:
    - ***Homoscedasticity*** of residuals (error terms),
    - ***Normal distribution*** of residuals
    - ***No perfect multicollinearity*** between variables.
    - ***Mean of Residuals*** are zero
    - ***No Correlation*** of error terms

  The numerical and/or graphical validations (& brief explanations) for assumptions 1,2 & 4 have been provided in the code. The others have been avoided for sake of simplicity but more information about the assumptions can be found in the link provided in the references.

Choosing the Right Variables

Keep in mind that, since Sarah is a female, we would like to find out the mean pass rate for females at Bury. But to begin our analysis, we would like to find out the possible variables that can potentially have an impact on female pass rates.

In order to assess this, we need to first inspect the variables available in our dataset:

```
#Inspecting names to look for explanatory variables
names (df_bury)
##  [1] "Test_Centre"           "Age"
##  [3] "Male_Tests_Conducted"   "Male_Passes"
##  [5] "Male_Pass_Rate"         "Female_Tests_Conducted"
##  [7] "Female_Passes"          "Female_Pass_Rate"
##  [9] "Total_Tests_Conducted"  "Total_Passes"
## [11] "Pass_Rate"              "Year"
```

Three intuitive choices that may have a causal impact on female pass rates could be: "Test Centre", "Age" and "Year". It is hard to tell whether the other variables have an impact merely by inspection.
Notice that Age is a ***numeric*** variable (it can take any value between negative & positive infinity) whereas Test Centre and Year are ***categorical*** variables (i.e. not numeric – they belong to a specific number of categories).
We can construct a few plots to decide which of Age and Year seem to affect female pass rates.

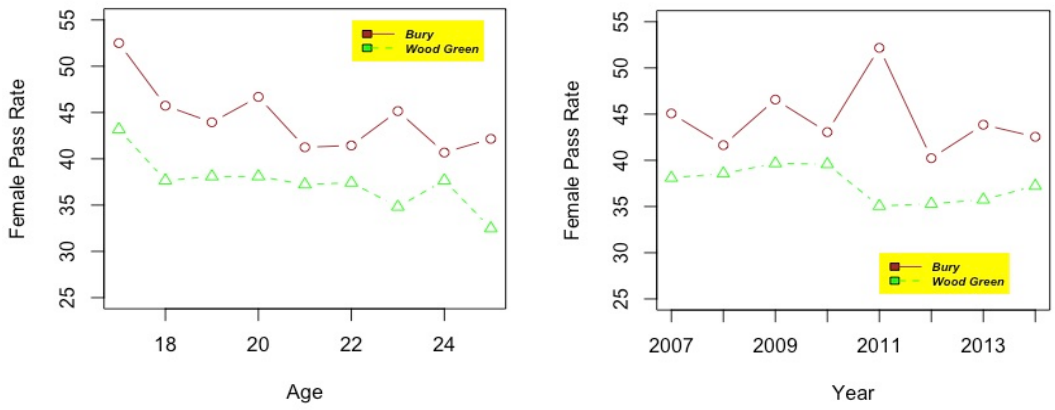

Figure 1 Relationship between Female Pass Rate vs Age & Year

As you can see, there is an interesting pattern between **pass rates** & **Age**; the older an individual is, they seem to have a lesser chance of passing in both Bury and Wood Green. This would (somewhat) validate my assumption that Sarah cannot wait another year to do the test because the older she gets, her chances of passing seem to falter.
For the relationship between **pass rate** vs **year**, other than the spike between 2010 & 2011, there doesn't seem to be any obvious pattern between the time period in consideration. Thus, I will ignore year as a predictor in my linear regression model below.
Taking the data at face value, we will calculate the mean, assuming that Female pass rates in Bury and Wood Green depend only on Age & Test Centre:

```
Bury.Mean = mean(df_bury$Female_Pass_Rate[df_bury$Age == 22])
WoodGreen.Mean = mean(df_wgreen$Female_Pass_Rate[df_wgreen$Age == 22])
print(Bury.Mean)
```

```
## [1] 41.43886
print(WoodGreen.Mean)
## [1] 37.42147
```

As we would have guessed from the graph above, Sarah's expected pass rate at Bury would be higher than at Wood Green by around 4 %.
Is there a way to test whether this observation is statistically sound? This is the subject of the next few sections.

How Accurate is the Mean?

In order to statistically validate the claim that mean female pass rates are higher in Bury, I will begin by using the method of ***Linear Regression***.

In simple terms, Linear Regression is "a **linear** approach for modelling the relationship between a scalar dependent variable y and one or more explanatory variables (or independent variables) denoted X. The case of one explanatory variable is called **simple linear regression**" (Source: Wikipedia). Since we are regressing Female Pass Rate on more than one independent variable (i.e. Test Centre & Age), this is called ***multiple linear regression.***

The model used for multiple linear regression in the case of 2 variables (Age & Test Centre) is as follows:

$$Y \simeq \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

β_0 is the ***intercept*** term- that is, the *expected/ average value of Y when the X terms are 0.*
The $\beta's$ are the ***slope*** terms – for example, β_1 is the *expected value of Y when* X_1 *goes up by one unit.*
The term ε is called the irreducible-error term/ **residual** – the true relationship between Y and X_i's may not be linear, or there may be other variables affecting Y which haven't been accounted for. The ε accounts for all these discrepancies. Visually, the residual can be thought of as the gap between the **actual** value of Y (denoted by a point in the graph) and the **predicted** value from our regression line (perpendicular to the point).

The following figure fits a ***line-of-best-fit*** to the relationship between Female pass rates vs Year. A best-fit line is "a straight line that best represents the data on a (scatter) plot" (Source: Wikipedia):

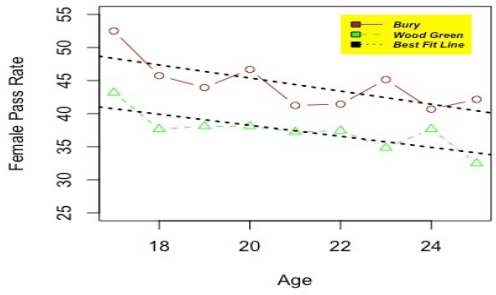

Figure 2 Best Fit Line showing the accuracy of a linear fit to the data.

The above representation also provides more motivation for performing linear regression as the line seems to fit the data reasonably well. Furthermore, there doesn't seem to be any extreme observations (outliers), where the predicted values from the regression line seem to be far away from the actual values (denoted by the points).

Interpreting a Qualitative Variable

In our model, the "Test Centre" variable is ***Qualitative/Categorical*** – this means that it isn't numeric and belongs to a certain number of categories. There are 3 categories that it belongs to: Bolton, Bury (Manchester) and Wood Green (London).
An important note to make here is that since there is no data for Bury (Manchester) in 2011-12, I am taking the data for Bolton, the next closest Test Centre to Bury. Thus, I will rename all the rows named "Bolton" to "Bury (Manchester)".
To incorporate this in our regression model, we need to create a ***dummy variable*** that takes two possible values:

$$wgreen_i = \begin{cases} 1 \; if \; the \; ith \; person \; \; is \; in \; Wood \; Green \; (London) \\ 0 \; if \; the \; ith \; person \; \; is \; in \; Bury \; (Manchester) \end{cases}$$

This will result in the model:

$$y_i = \beta_0 + \beta_1 age_i + \beta_2 wgreen_i + \varepsilon_i$$
$$= \begin{cases} \beta_0 + \beta_1 + \beta_2 + \varepsilon_i \; if \; the \; ith \; person \; \; is \; in \; Wood \; Green \; (London) \\ \beta_0 + \beta_1 + \varepsilon_i \; if \; the \; ith \; person \; \; is \; in \; Bury \; (Manchester)) \end{cases}$$

The model with no dummy variables is called the ***baseline model.***
Note that the decision we make to code Bury (Manchester) as 0 and Wood Green as 1 is completely arbitrary and doesn't change the estimate of the regression fit.
Since we are ***estimating*** the above model, we will write it as follows:

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 age_i + \hat{\beta}_2 wgreen_i$$

Note that the error term ε has been cancelled because we assume its expected value to be zero (i.e. the sum of all the positive & negative errors from estimating through a best-fit line should sum to zero).

Weaknesses of the Model

It is important to highlight that, although our model seems to fit the data well, it is prone to some weaknesses:

- **Other Significant Variables**: that could have impacted female pass which hasn't been accounted for. Example :- Gender & Confidence level(experience) of Sarah.
- **Small Data Set:** Contains only 144 total observations and may give us a misleading conclusion than if a larger dataset was used. With a larger dataset, it is possible that the relationship between female pass rates vs the predictors Age/Year/Test Centre could be very different.
- **Leverage:** There may be extreme observations for the predictors (X values) that can cause misleading results in the regression fit. This has been visually demonstrated in the code. An improvement could be to remove the observations with high leverage.

Regression Model in R

We have now come to the moment of truth where our model is about to inform Sarah which location would be best for her to do the test. We are going to carry out the above model in R using the code as follows:

```
#Make a new data frame combining Bury and Wood Green.
df_bury_wgreen <- rbind(df_bury, df_wgreen)
#Rename the Bolton Factor to Bury (Manchester) to be consistent with our initial assu
mption -
#Since there is no data for Bury (Manchester) in 2011-12, I am taking the data for Bolto
n, which is the next
#closest Test Centre to Bury of approximately 6 miles.
df_bury_wgreen$Test_Centre <- c(rep("Bury (Manchester)",72), rep("Wood Green (Lond
on)", 72))
# Convert the test centre variable from character to factor in order to display contrasts
(shown below)
df_bury_wgreen$Test_Centre <- as.factor(df_bury_wgreen$Test_Centre)
# Run a linear Regression model on Female Pass Rate against Test Centre and Age fro
m the df_bury_wgreen dataset.
# Note that the command to run linear regression in R is lm(y ~ x, data = ) where lm st
ands for linear model.
lm_bury_wgreen <- lm(Female_Pass_Rate ~ Age + Test_Centre , data = df_bury_wgreen)
# Return the summary statistics of our linear model
summary(lm_bury_wgreen)
##
## Call:
## lm(formula = Female_Pass_Rate ~ Age + Test_Centre, data = df_bury_wgreen)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -11.8533  -3.4499  -0.4407   2.7066  13.6171
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  63.5433     3.4015  18.681   < 2e-16 ***
## Age                          -0.9119     0.1596  -5.714   6.30e-08 ***
## Wood Green (London)  -6.9871     0.8241  -8.479     2.77e-14 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.944 on 141 degrees of freedom
## Multiple R-squared:  0.4258, Adjusted R-squared:  0.4176
## F-statistic: 52.27 on 2 and 141 DF,  p-value: < 2.2e-16

# Display the coding scheme used for the dummy variables
contrasts(df_bury_wgreen$Test_Centre)
##                  Wood Green (London)
## Bury (Manchester)                   0
## Wood Green (London)                 1
```

I will only attempt to explain the relevant statistics in order to understand the model. If you need a more in-depth explanation, type in the following command into R:
?summary.lm

Under the Estimate column, we have the value for the intercept ($\hat{\beta}_0$), Age ($\hat{\beta}_1$), Bury Manchester ($\hat{\beta}_2$) and Wood Green (London) as $\hat{\beta}_3$. The stars at the end of each row tell us how significant/ important each variable is in estimating Female Pass Rates, with more stars indicating more significance.
The model would then be:

$$female\_pass\_rate_i = \hat{\beta}_0 + \hat{\beta}_1 age_i + \hat{\beta}_2 test\_centre_i$$

With that brief explanation, we can now proceed to the calculations:
The Expected female pass rate for a 22 year old at Wood Green would be:

```
63.5433 - 0.9119*(22) - 6.9871
## [1] 36.4944
```

The Expected female pass rate at Bury (Manchester) is:

```
63.5433 - 0.9119*(22)
## [1] 43.4815
```

Hence, Sarah's chances of passing would be approximately 7 % better if she chose Bury as her test center. Voila!

Statistical Accuracy of The Model

So far, we have shown (through linear regression) that our initial finding of the mean pass rate at Bury being higher than at Wood green holds. But is it statistically accurate? Let us use the F-test & p-values to try & prove it.

You might be thinking, what is an F-test & what are p-values?
Simply put, an F test is used to judge if at least one of the predictor variables are significant/ important in affecting the dependant variable or not. This is measured by the coefficients of the predictors (β_i's). It is based on two hypotheses, the ***null*** & ***alternative***:

H_o*(null)* : All β_i's are 0
H_1*(alternative)* : At least 1 β_i not equal 0.

The p-value is basically the risk of rejecting the null hypothesis in favour of the alternative: a small p-value would indicate a low risk of rejecting the null and thus, we would reject it. Similarly, a high p-value would indicate a large risk of rejecting the null, and so we do not reject it. Note that for statistical purposes we NEVER say that we "accept" the null in the latter case.

For our model above, we can see that the F-statistic is very high at 52.21 & the p-value is extremely small at 2.2*10^(-16). That's a lot of zeros! This means we should reject the null and conclude that at least one predictor affects the female pass rate. If we were to look at the individual p-values for the predictors, we find that even they are extremely small which means that both predictors are important in predicting female pass rates.

So Sarah, I think it would be very wise if you chose to go to the test Centre at Bury to get your license.

References

- Introduction to Statistical Learning: http://www-bcf.usc.edu/~gareth/ISL/

- Using Linear Regression to Predict Energy Output of a Power Plant: https://www.r-bloggers.com/using-linear-regression-to-predict-energy-output-of-a-power-plant/

- Image credits: https://surety1.com/surety-bond-driving-schools/

- Adding legend to an R plot: http://www.sthda.com/english/wiki/add-legends-to-plots-in-r-software-the-easiest-way

- Regression Analysis (Interpreting the Constant): http://blog.minitab.com/blog/adventures-in-statistics-2/regression-analysis-how-to-interpret-the-constant-y-intercept

- Linear Regression Assumptions: http://r-statistics.co/Assumptions-of-Linear-Regression.html