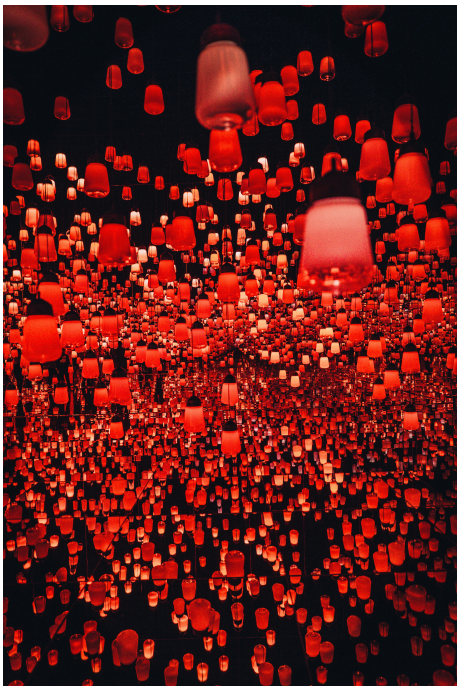

ETC513 Assignment 3: Comparison of Energy and Pollution by Country

Shaohu Chen
Master of Business Analytics(In Progress)

Qian Duan
Master of Business Analytics(In Progress)

Tina Tsou
Master of Business Analytics(In Progress)

Report for
Australian Government COVID19

3 June 2021

Our consultancy
add names &
add names

(03) 9905 2478
questions@company.com

ABN: 12 377 614 630

Introduction

Step 1

Step 2

Step 3

In the following section, we will be analyzing the relationship between *Booking Type* and *Exam Result*.

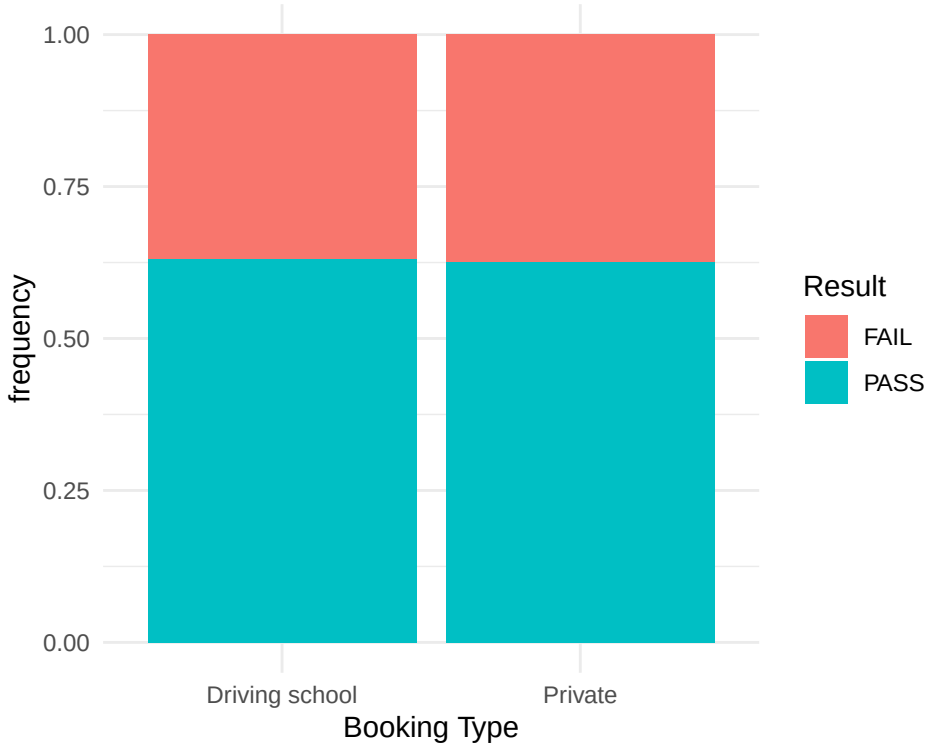

Figure 1: *Frequency Plot between Booking Type and Exam Result*

The frequency plot, Figure 1, between *Booking Type* and *Exam Result* shows that the percentages of people who passed the exam are similar for both driving school and private.

Since the response variable and predictor variable are categorical variables, they will have to be converted into dummy variables(0 & 1). Then, following Alice (2018), I ran a logistic regression to analyze their relationship. The following is the formula for the regression *logmodel*:

$Y \sim B(p)$, $log(\frac{p}{1-p}) = \beta_0 + \beta_1 X + \epsilon$

- β_0 is the intercept.

- β_1 is the coefficient of *Booking Type_Private*
- X is *Booking Type_Private* taking values 0 or 1

Table 1: *Regression Result for logmodel*

	Dependent Variable:
term	'Exam Result_PASS'
(Intercept)	0.533407196218835 ***
'Booking Type_Private'	-0.0179325664251033 *
nobs	337084
AIC	444926.751217685
logLik	-222461.375608843

Note:
makecell[l]p-value: '+' 0.05 - 0.1; '*' 0.01 - 0.05; '**' 0.001 - 0.01; '***' 0 - 0.001

Table 1 shows the regression summary. *Booking Type_Private* has p-value close to 0 which means it is statistically significant. Due to the variable being a dummy variable relative to booking type driving school, the coefficient indicates that *Booking Type_Private* affects the passing of an exam negatively compared to *Booking Type_Driving School*. Private booking reduces the log odds by 0.061.

ANOVA test, table 2, on the *logmodel* analyzes the table of deviance which shows how well the x variable is doing in comparison to the null model. Here we can see that the drop in deviance is quite small despite having low p-value.

Next, we test the fit of the model by looking at the receiver operating characteristic (ROC) curve.

Figure 2 shows the ROC curve of the *logmodel*. It is basically a 45 degree diagonal line which indicates the model has no discrimination ability.

Area under the curve "... gives the probability that the model correctly ranks such pairs of observations" Bartlett (2014). The area under the curve for this model is 0.5022363. In conclusion, the predictor just makes random guesses.

We try to improve the model by adding more variables to the function:

Table 2: *Anova for logmodel*

	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)
NULL	NA	NA	337083	444929.1	NA
'Booking Type_Private'	1	6.315064	337082	444922.8	0.0119716

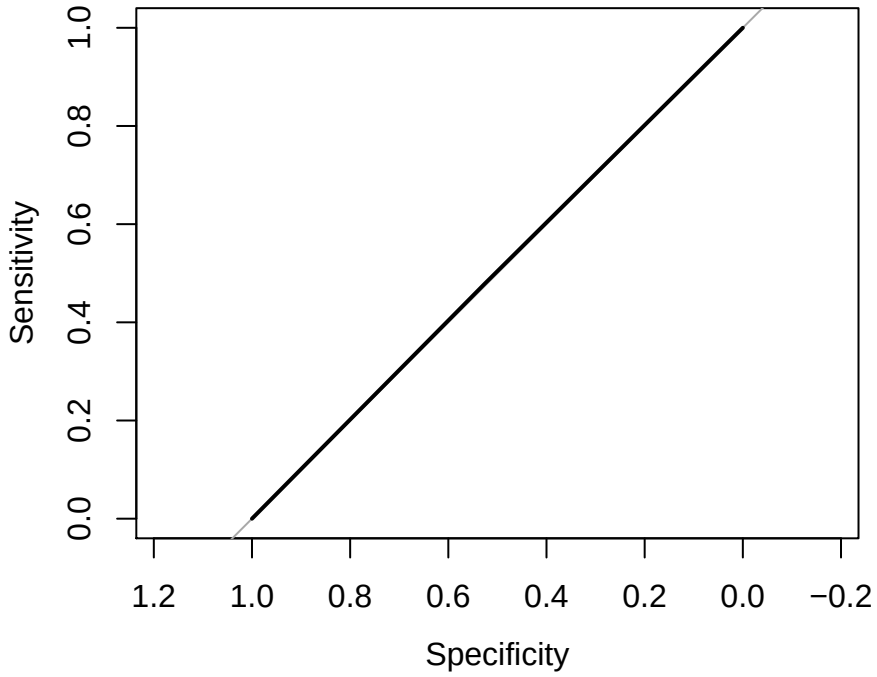

Figure 2: *ROC Cuve of logmodel1*

$Y \sim B(p)$, $log(\frac{p}{1-p}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$

- X_2 is *Number of Examinations* taken by each examinee.

Table 3: *Regression Result for logmodel*

	Dependent Variable:	
term	'Exam Result_PASS' (logmodel2)	'Exam Result_PASS'(logmodel)
(Intercept)	0.522573474570028 ***	0.533407196218835 ***
'Booking Type_Private'	-0.0157422192951367 *	-0.0179325664251033 *
'Number of Examinations'	0.00169271410738006 ***	
nobs	337084	337084
AIC	444885.718925042	444926.751217685
logLik	-222439.859462521	222461.375608843

Note:
makecell[l]p-value: '+' 0.05 - 0.1; '*' 0.01 - 0.05; '**' 0.001 - 0.01; '***' 0 - 0.001

Table 3 shows the two regression summary side-by-side. The regression with *Number of Examinations* has AIC of 333404. It is slightly lower than the AIC of the previous regression which was 333483. Thus, in comparison, having this one extra variable improved the function significantly (statistically).

However, a simple calculation of area under ROC curve for logmodel2,0.4859687,indicates that the model is even worse than the first.

Conclusion

In conclusion, we've shown that higher pass rate in certain districts is not always an absolute reflection on whether the district has better driving program. Rather, it is an outcome of locations with lower examinees in general. For locations with more examinees, there would be more variations in their outcome thus more fails.

Next, we shown that automatic cars have the lowest pass rate overall, and that motorcycle (over 250cc) has the highest pass rate. Older people (66 and above) also tend to fail their vehicle tests more. But ultimately pass rate for each vehicle type and majority of the age group is over 50%.

Last but not least, although, there is statistical relationship between the booking type and the exam outcome, the affect is pretty small. Furthermore, the current variables are inadequate in creating a good model to predict the outcome.

This is also a shortcoming with the data we currently have. Because it contains very limited variables, it is hard to create a better fit model that can predict the outcome accurately.

References

Alice, M (2018). *How to Perform a Logistic Regression in R*. https://datascienceplus.com/perform-logistic-regression-in-r/.

Bartlett, J (2014). *Area under the ROC curve – assessing discrimination in logistic regression*. https://thestatsgeek.com/2014/05/05/area-under-the-roc-curve-assessing-discrimination-in-logistic-regression/.