Regression Analysis of Road Traffic Collisions, Accidental Dwelling Fires and Youth Unemployment

1. Road Traffic Collisions

Limitation of Current Model

The Future Melting Pot's (TFMP) pilot study into Road Traffic Collisions' (RTC) analyzed data, through the use of graphs, to show trends between relevant variables. The method's limitation is that graphs can only show the relationship between two variables, with the lack of coefficient that can be used to understand the numerical relationship between factors. Moreover, the reason a RTC has occurred may be contributed to the joint effect of two or more variables, such as cultural identity, road condition, and weather etc. Thus, a regression model is a more effective test what are the influential factors that causing RTCs in the three study sites.

Purpose

The main purpose of the regression model includes:

- To test the effect of; gender, age, ethnicity, household disposable income (per month) and self-expression on RTCs.
- Identify the significant factors in the model to develop better targeted RTC prevention methods.

Methodology

In order to estimate the probability of RTCs in the three study sites,

I will use the Probit and Logit regression models for the analysis. The results of these models will inform my conclusion.

Logit Model: F (Z) =$\frac{1}{1+exp^{-Z}}$

Probit Model: $F(Z) = \int_{-\infty}^{Z} \Phi(u)dx$

Where Z = Road Traffic Collisions

= β0 + β1 Male + β2 Age +β3 ethnicity + β4 disposable income (per month) + β5 self-expression +β6 Location + Ui

Noted:

The ethnicity have been grouped into 5 categories, thus, a total of 4 variables (N-

1) should be included in the model as independent variables.

Self-expression need to be categorized into different level by further interviews and/or data collection processes so as to deepen the understanding of how cars and driving act as representations of social mobility.

Locations constitute Hodge Hill, Sheldon, Solihull.

Both Logistic and Probabilistic regression are able to estimate the probabilities of an outcome where the events are coded as binary variables. Binary variables can also be modeled into Ordinary Least Square (OLS) using linear probability model (LPM) (Menard 1995, p 6). Compared to LPM, Probit and Logit are effective models that can produce fitted values falling inside the acceptable probability range by 0<=prob<=1 with binary dependent variable of interest, while the predict value ($\hat{y}$) may fall outside the 0 to 1 range using LPM. Furthermore, the assumption of heteroscedasticity is more likely violated with a binary dependent variable in the model, especially when the probability of the dependent event varies widely (Pohlman and Leitner 2003). As a result, Probit and Logit models are the better choice for evaluating the probability of RTC for an individual in the West Midlands. The criteria for selecting between Probit and Logit models is the one that with smaller gap between observation and expectation value.

Data

The data for the regression analysis is cross-sectional and it comes from the record of West Midlands Fire Service/Police/Local Authorities. The statistic is drawn from three Birmingham electoral wards, Washwood Health, Sheldon and Elmdon in Solihull, with 513 observations in total. The dependent variable is road traffic collisions while the independent variables include gender, age, ethnicity, disposable income and self-expression.

To model the probability of being in an accident, information on both those that has crashes and those that don't is needed.
With the lack of data for people who are not involved in RTCs for the three study sites, further testing of influential factors that correlated to RTCs is limited.

Recommendations

My recommendation for improving the RTC prediction is to collect the data for every individual in the three study sites in terms of age, gender, ethnicity, disposable income and self-expression. With sufficient data, we are able to identify the significant factors that cause RTCs by checking the p-value. Moreover, the regression model may contain many variables that operate independently, or in connection with one another, to explain variation in the dependent variable (RTCs). Lastly, the coefficient of each variable can suggest the incremental

probability of RTCs while holding other related independent variables constant.

For example, in the Probit Regression Model, if the coefficient of male is 0.8 and significant (if p-value less than 0.05), that means that the z-score of Pr (A Road Traffic Collision Happening for An Individual) is 80% higher for a male compared with a female in the three study sites.

Regression Model of RTCs Severity

Based on the current collected data (file name: Breakdown Analysis RTC Data for The Future Melting Pot), a regression model can be built to explore the factors that correlated to impact the severity of an RTC. The dependent variable (outcome variable of interest) is the severity of the RTC and the independent variables are gender, age, ethnicity, locations. Based on the data given, severities are categorized into four levels, which are slight, less serious, moderately serious and very serious. I coded "less serious", "moderately serious" and "very serious" as the event of happening (Y=1) and "Slight" as the event of not happening. Hence, we are able to build a binary choice model that measures the probability of outcome that falls into "serious" with certain independent variables as input.

During the October 2015 to October 2016 period, there are 190 observations in total for RTCs. In Ethnicity, there are 44 instances of unknown/unrecorded data and it's rational to code them as "Other" to take them into further analysis. In addition, I removed 4 observations in the dataset which are lack of age information.

Results

Table I – Logit & Probit Model of Severity of RTCs by Gender, Age, Ethnicity and Location

Variables	Logit	Probit	P Value
Male	0.830	0.449	**> 0.5**
	(0.467)	(0.245)	
Age	0.006	0.003	**> 0.5**
	(0.010)	(0.005)	
Ethnicity	-0.057	-0.034	**> 0.5**
	(0.178)	(-0.097)	
Location	-0.287	-0.170	**> 0.5**
	(0.418)	(0.233)	
Constant	-0.287	-1.216	< 0.5
	(0.723)	(0.380)	

Psuedo R2	0.029	0.029	
N=	186	186	

In table 1, the results show that none of the independent variables are significant to the dependent variable with its p-value greater than 0.05. Gender, age, ethnicity and the region that people lived have therefore nothing to do with the severity of the RTC. Not surprisingly, the independent variables in the model are not correlated to causing either a serious RTC or a slight one. Consistent with the case-control studies in Iran and Hong Kong, severe injury of road traffic crashes were found to be associated with rainy weather, night time, speed of driving (Majdzadeh, 2008). That's most likely the reason why gender, age, ethnicity and the road condition on the study areas do not showing the effect of severity of RTCs. In addition, since the sample size is relatively small in measuring the severity level of RTCs for a city, including the data for a longer period of time (eg. 5 years) or/and observations from more study sites are recommended.

2. Predicting the ADF

To analyze the factors that correlated to ADFs (Accidental Dwelling Fires), data collection methodologies are important to identify the required variables for the construction, verification, validation and testing of the forecasting models (Taylan and Demirbas, 2016). Since we are more interested in identifying socio-economic and demographic factors which are associated with higher rates of dwelling fires, the following data sets are recommended to be recorded.

- Disposal income per month (after tax) for household
- Ethnicity
- Age
- Gender
- Education

152 ADFs were recorded across the three study sites between January 2013 and December 2015. However, only general information of victims such as the, total number of oven related incidents and number of electrical related fires etc. have been collected. Based on the data provided, we can't identify which factors have a significant impact on an ADF and which are not. By collecting more data relating to the characteristics of the victims it would allow for more comprehensive and effective quantitative predictions.. As a result, rational recommendations can be further suggested so as to better minimize the potential probability of an ADF caused by the identified significant variables.

By Including the Disposal income field, in the regression model is rational it

suggests the house quality and facilities of the individual. People with high income can afford a more expensive and better quality property. Ethnicity, age and gender could suggest the eating habits and ways of cooking within a specific group of people. Individuals with a higher level of education may indicate that he/she has a safer living habit or better knowledge in preventing ADFs. After identifying the significant variables of the model, it's worth to analysis them deeper with socio-economic and demographic effects or/and motivations.

3. Youth Unemployment

Despite youth unemployment rate decreasing in Washwood Health, the numbers were higher than the other two study sites, and almost twice the national average. Knowledge of the personal, regional and family background characteristics of unemployed youth is crucial to enable government to design effective policy. A better understanding of unemployment trend over the past 5-10 years and factors that correlated to the rate could serve as future indicators of fluctuation in the unemployment rate. Factors that related to youth unemployment can be tested by statistical analysis.

Data & Variables

The government defines unemployment as being, People who are aged 16 or over without work and available to start work in the next two weeks, if they have been either actively seeking work in the past four weeks or waiting to start a new job they have already obtained (Ons.gov.uk, 2018). Based on the first phase of the research, we have been analyzed the factors such as the ability to drive, family relationship history, the history of drug, criminal record etc. To provide a more accurate result, I suggest collecting further data on all youth between age 18-24 in the three study sites. Moreover, data can be recorded and categorized in three sets that influence unemployment outcomes, which are personal characteristics (such as low education level or no formal qualifications, low language ability, and attendance to large schools with lower academic achievements), family characteristics (such as presence of unemployed parents or single-parent families), and regional characteristics (such as limited access to public transportation) (Viitanen, 1999).

Methodology

A statistical analysis can be conducted to determine if any relationship between the unemployment and six selected variables existed in order to identify trends in the unemployment and to predict to unemployment rate for a given year by interpreting the variable coefficients.

Logit Model: $F(Z) = \frac{1}{1+exp^{-Z}}$

Probit Model: $F(Z) = \int_{-\infty}^{Z} \Phi(u)dx$

Where Z = Youth unemployment
= β0 + β1Personal Characteristics+ β2 Family Characteristics +β3 Regional Characteristics + Ui

Independent variables should be further coded in the regression model based on the availability of data in next stage. With sufficient data, regression result can be interpreted and further analyzed in the next stage.

Summary

In the pilot study, graphs have been used to show the relationships between two variables. The lack of a coefficient that can be used to understand the numerical relationship between factors entails the necessity of a regression model. To predict the probability of RTC, ADF and youth unemployment, further data collection is recommended in the three study sites. With sufficient data, we can identify significant correlating factors. As a result, the follow-up regression models in the next stage will provide the policy makers with a reliable and useful resource when addressing societal issues.

References

The Ohio Journal of Science. (2003). *A Comparison of Ordinary Least Squares and Logistic Regression.* [online] Available at: https://pdfs.semanticscholar.org/5a20/ff2760311af589617ba1b82192aa42de4e08.pdf

Majdzadeh R, Khalagi K, Naraghi K, Motevalian A, Eshraghian MR. Determinants of traffic injuries in drivers and motorcyclists involved in an accident. Accid Anal Prev. 2008;40((1)):17–23.

Taylan, O. and Demirbas, A. (2016). Forecasting and analysis of energy consumption for transportation in the Kingdom of Saudi Arabia. *Energy Sources, Part B: Economics, Planning, and Policy*, 11(12), pp.1150-1157.

Ons.gov.uk. (2018). *Unemployment - Office for National Statistics*. [online] Available at: https://www.ons.gov.uk/employmentandlabourmarket/peoplenotinwork/unemployment [Accessed 27 Apr. 2018].

Viitanen, T. (1999). *Estimating the Probability of Youth Unemployment.* [online] Available at:

http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.23.1623&rep=rep1&type=pdf

Appendices

Appendix 1 – DO FILE for STATA

```
use "C:\Users\chenk11\Desktop\TFMP RTC DATA.dta"
gen male = 0
replace male = 1 if sex == "MALE"
gen ethnicity1 = 0
replace ethnicity1 = 1 if ethnicity == "ASIAN"
replace ethnicity1 = 2 if ethnicity == "BLACK"
replace ethnicity1 = 3 if ethnicity == "WHITE - NORTH EUROPEAN"
gen Severity = 0
replace Severity=1 if severityofcasualty == "MODERATELY SERIOUS"
replace Severity=1 if severityofcasualty == "VERY SERIOUS"
replace Severity=1 if severityofcasualty == "LESS SERIOUS"
gen location1 = 0
replace location1 =1 if location == "Hodge Hill"
replace location1 = 2 if location == "Sheldon"
des age
destring age, replace
destring age, replace force
summarize male age ethnicity1 Severity location1
regress Severity male age ethnicity1 location1, r
probit Severity male age ethnicity1 location1
logit Severity male age ethnicity1 location1
logit Severity male age ethnicity1 location1, or
probit Severity male age ethnicity1 location1, r
probit Severity male age ethnicity1 location1
margins, dydx(*)
quietly regress Severity male age ethnicity1 location1, r
predict ols
quietly probit Severity male age ethnicity1 location1, r
predict probit
estat gof, group (10) table
quietly logit Severity male age ethnicity1 location1, r
predict logit
estat gof, group (10) table
sum Severity ols probit logit
```

Appendix 2 – Stata output of Probit Regression

```
. probit Severity male age ethnicity1 location1

Iteration 0:   log likelihood =  -78.79714
Iteration 1:   log likelihood = -76.511845
Iteration 2:   log likelihood = -76.497227
Iteration 3:   log likelihood = -76.497226

Probit regression                               Number of obs     =        186
                                                LR chi2(4)        =       4.60
                                                Prob > chi2       =     0.3309
Log likelihood = -76.497226                     Pseudo R2         =     0.0292
```

Severity	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
male	.4489128	.2450554	1.83	0.067	-.0313869	.9292126
age	.002978	.0054537	0.55	0.585	-.007711	.0136669
ethnicity1	-.0338447	.097477	-0.35	0.728	-.2248962	.1572068
location1	-.1704792	.2331766	-0.73	0.465	-.6274969	.2865385
_cons	-1.215885	.380378	-3.20	0.001	-1.961412	-.4703575

Appendix 3 – Stata Output of Logistic Regression

```
. logit Severity male age ethnicity1 location1

Iteration 0:   log likelihood =  -78.79714
Iteration 1:   log likelihood = -76.571412
Iteration 2:   log likelihood = -76.509711
Iteration 3:   log likelihood =  -76.50963
Iteration 4:   log likelihood =  -76.50963

Logistic regression                             Number of obs     =        186
                                                LR chi2(4)        =       4.58
                                                Prob > chi2       =     0.3337
Log likelihood =  -76.50963                     Pseudo R2         =     0.0290
```

Severity	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
male	.8295999	.4671351	1.78	0.076	-.0859682	1.745168
age	.0057685	.0100643	0.57	0.567	-.013957	.0254941
ethnicity1	-.0569823	.1781003	-0.32	0.749	-.4060525	.292088
location1	-.2872635	.4175963	-0.69	0.492	-1.105737	.5312103
_cons	-2.122249	.7232345	-2.93	0.003	-3.539763	-.7047352