What explains domestic violence rates in NSW?

Masud Jubaier

Introduction

Domestic violence (DV) is dark stain in modern society. DV can be defined as any form of mistreatment including physical, phycological, financial, sexual take place within domestic circumstances perpetrated by member of intimate relationship. Any member of household can be victim of DV, while women and children are most vulnerable to this crime. The World Health Origination(WHO) has described the violence against women is a global problem of public health which need immediate attention[1]. Australia is not Immune to this crime, especially NSW suffer with this crime heavily. Hance it is very important the understand the nature of this crime and identify the factors having most influence. In this project we shall try find these factors and try to prove following hypothesis by implementing a linear model.

Null Hypothesis (H_0): There is no relationship between income, education and employment with DV.
Alternative Hypothesis (H_A): There is strong correlation between income, education and employment with DV.

Data and methodology

The dataset contains 3 CSV file. The 1^{st} CSV file "DV_NSW_by_LGA.csv" contains the count of DV crime in NSW LGA from Jan-99 to Dec-15. The 2^{nd} CSV file "NSW_LGA.csv" contain 2011 census data of NSW LGA. In the 2^{nd} CSV file there are around 8K features on 2011 NSW LGA census data. The features are in encoded form, where the description of the features is found in the 3^{rd} CSV file "labels.csv". At first we have performed some exploratory analysis to understand the dataset. We have generated some visualization to show the trend of the crime. From the visualization it is clear the trend is upward. To identify the features contributes most towards the crime we have implemented Forward stepwise Feature selection.

Forward stepwise Feature selection is a computationally efficient algorithm to select the best features that produces best outcome. The algorithm starts with a null model and add predictors one by one until all predictors are tried. At each step, the predictor that give the largest improvement on the model is added to the final predictor list. The steps of the algorithm are as follows.

1. Let, M_0 be a null model, which has no predictor.

2. For $k = 0, .. , p - 1$:

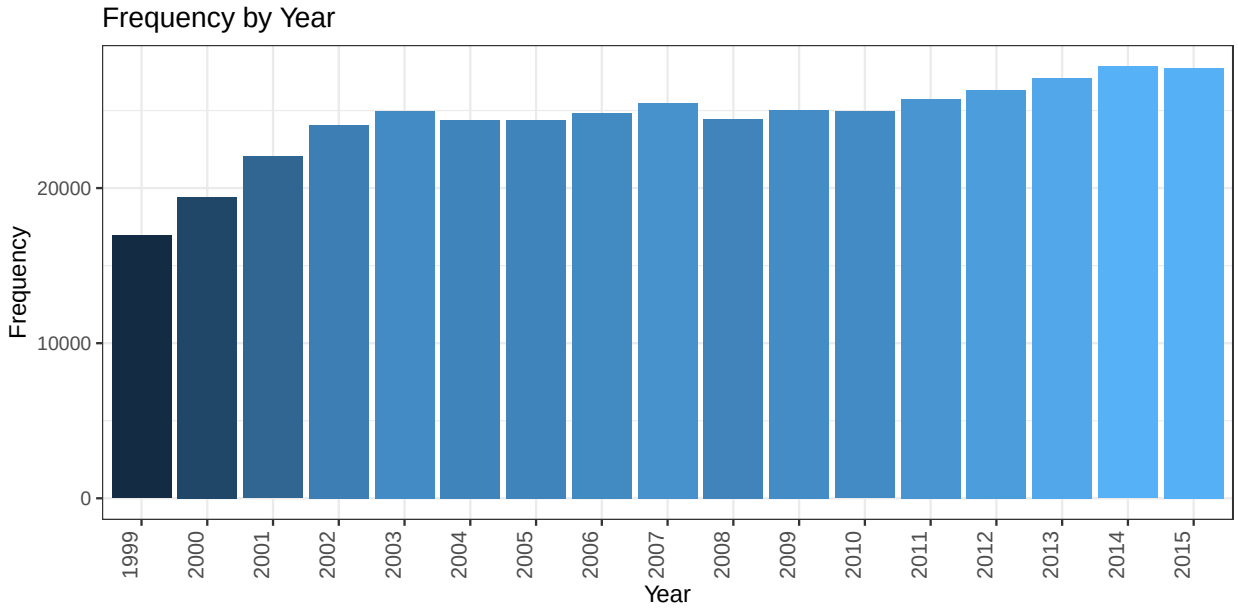
   (a) Consider all $p - k$ models that supplement the predictors in M_k with one added predictor.
   (b) Chose the *best* between these $p - k$ models, and name it M_{k+1}. Here *best* is considered as the model having minimum RSS or maximum R^2.

3. Select the best model from M_0, .., M_p using any one of these metric cross-validated prediction error or C_p (AIC) or (BIC) or adjusted R^2.

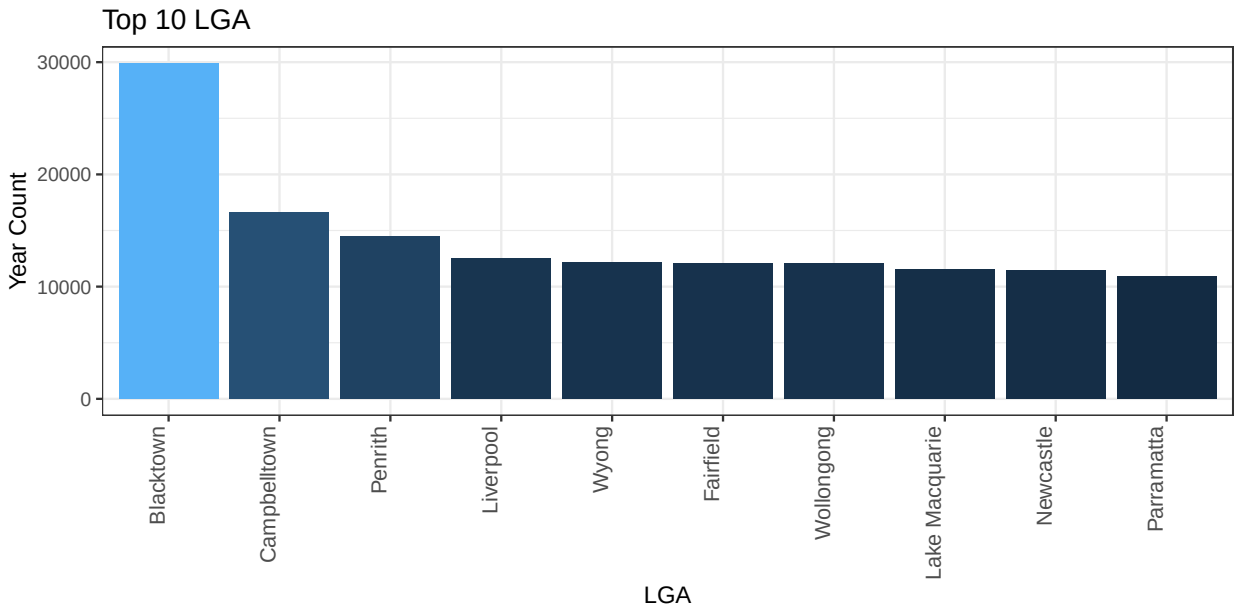
Using the features identified we shall build a linear model to predict DV.

Results and implications

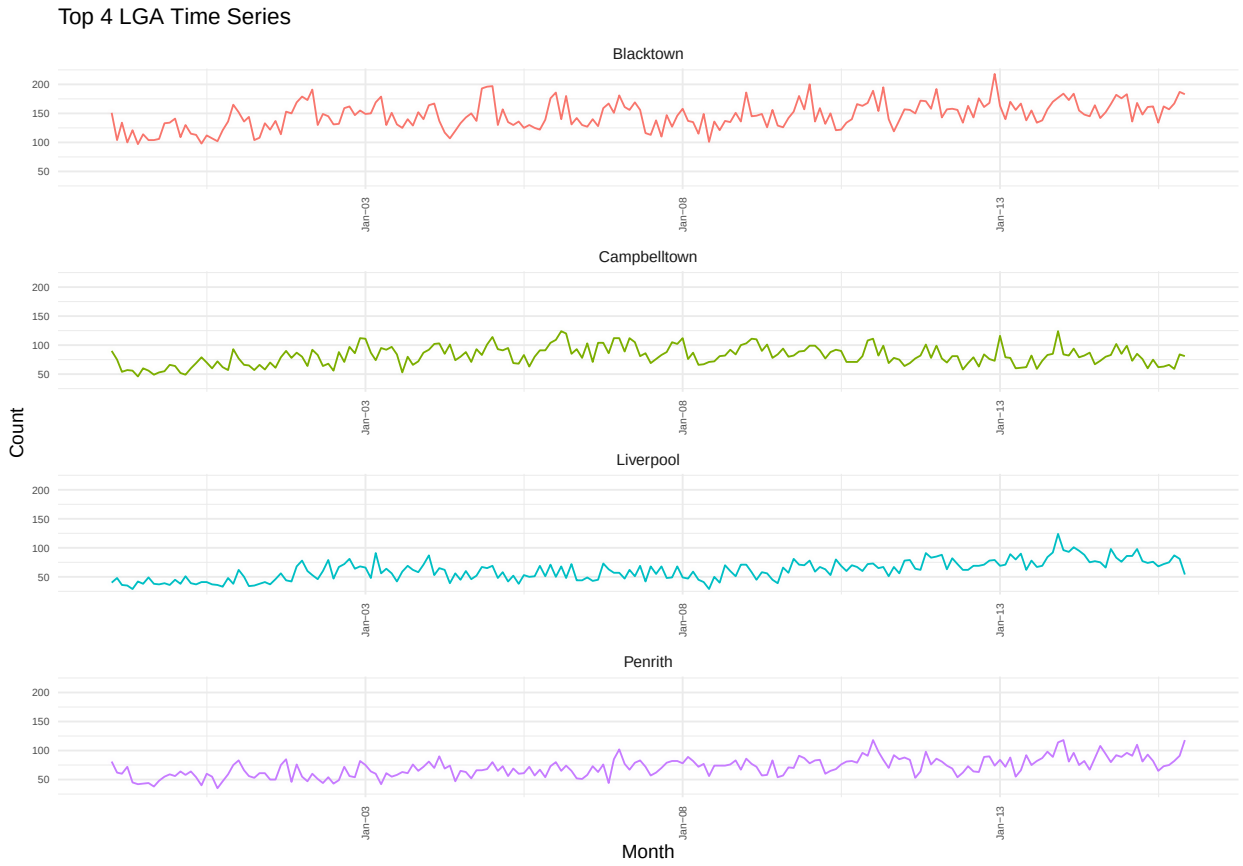
The plot below visualizes yearly frequencies of all across NSW. From the plot it is clear the crime rate has an upward trend.



The plot below visualizes the frequency of top 10 LGA in descending order. Blacktown LGA outcast all other LGA by very clear margin. The concern authority might need to pay extra attention on Blacktown.



The plot below visualizes the frequency timeseries of top 4 LGA. The timeseries plot shows normal distribution of the frequency, no sudden abrupt spike or drop of the frequency.



After applying the Forward stepwise Feature selection algorithm, the features selected by the algorithm are listed in the table below. From the feature list we can see, no features on income education and employment as our hypothesis are selected by the algorithm. Therefore, we are manually adding feature B115, B126, B5503 and B2847 in the predictor list covering these aspects. With this combined predictor list, we shall build a linear model.

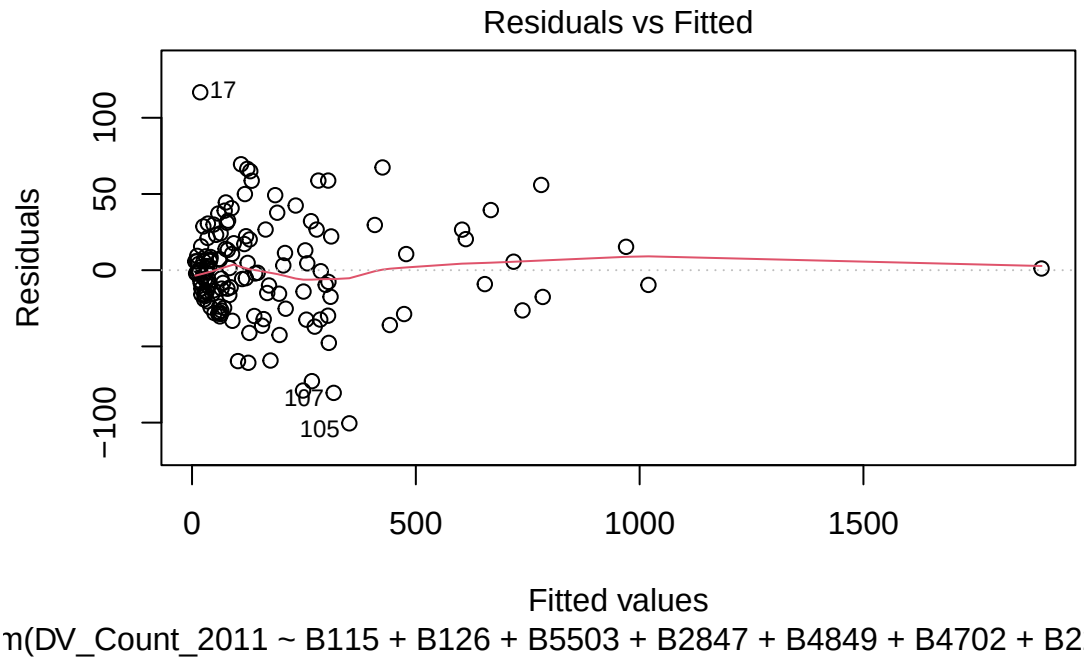
We are building a linear model with the selected features. The summary of the linear model as presented below.

```
##
## Call:
## lm(formula = DV_Count_2011 ~ B115 + B126 + B5503 + B2847 + B4849 +
##     B4702 + B2252 + B4258 + B2894 + B1758 + B3851 + B7452 + B2409 +
##     B1390, data = DV_NSW_2011_all)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -100.473  -17.105   -1.651   18.293  116.720
##
## Coefficients:
```

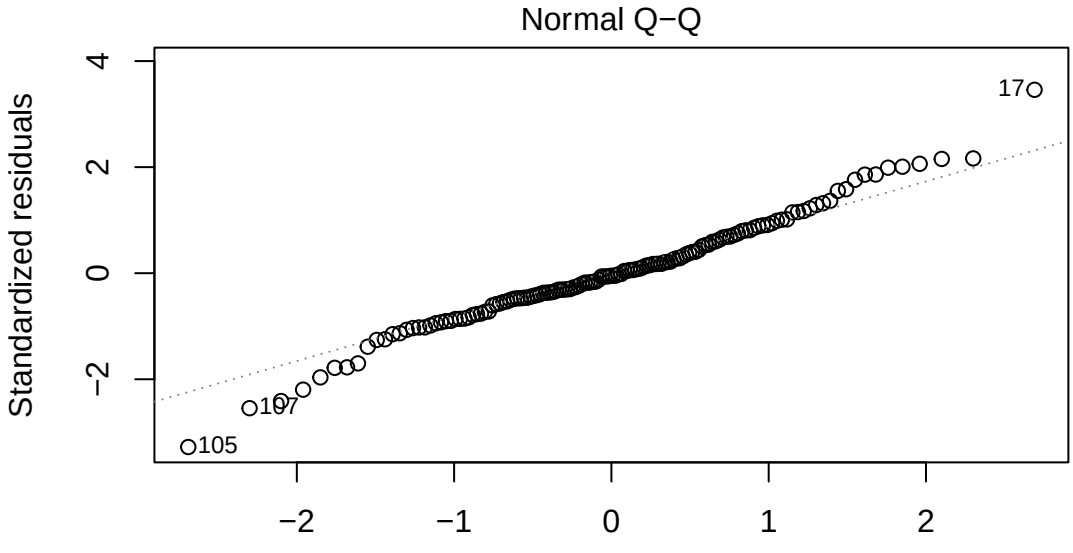
Sequential	Long
B1390	Cambodia Year of arrival 2011
B1758	Poland Year of arrival 2011
B2252	Dependent children aged 5 9 years female parent Language and proficiency in English not stated male parent Speaks English only
B2409	Dependent children aged 15 17 years female parent Total male parent Speaks other language and speaks English Proficiency in English not stated
B2894	Males Year 11 or equivalent Age 15 19 years
B3851	Females 15 19 years Did unpaid domestic work 30 hours or more
B4258	Persons 20 24 years Cared for Total
B4702	Age group of parent 25 29 years Number of children ever born Six or more children
B4849	One parent family with children under 15 and no dependent students and non dependent children Persons
B7452	Professional scientific and technical services Occupation Labourers

```
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 41.0157167 32.3057487   1.270 0.206582
## B115         0.0015837  0.0141017   0.112 0.910762
## B126        -0.0002338  0.0005186  -0.451 0.652966
## B5503       -0.6446707  0.7519112  -0.857 0.392878
## B2847        0.0306113  0.0207253   1.477 0.142191
## B4849        0.2614197  0.0394617   6.625 9.30e-10 ***
## B4702        7.7695686  1.6862621   4.608 9.90e-06 ***
## B2252        9.1174831  2.3752840   3.838 0.000196 ***
## B4258        0.1956317  0.0460197   4.251 4.13e-05 ***
## B2894       -0.8215502  0.1847910  -4.446 1.91e-05 ***
## B1758       11.3768055  3.1433709   3.619 0.000427 ***
## B3851        2.3630352  0.5996256   3.941 0.000134 ***
## B7452        0.8505166  0.3691961   2.304 0.022891 *
## B2409       -3.4327696  1.7094242  -2.008 0.046784 *
## B1390       -2.5787578  0.9509693  -2.712 0.007636 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 34.06 on 125 degrees of freedom
## Multiple R-squared:  0.9831, Adjusted R-squared:  0.9812
## F-statistic: 519.3 on 14 and 125 DF,  p-value: < 2.2e-16
```

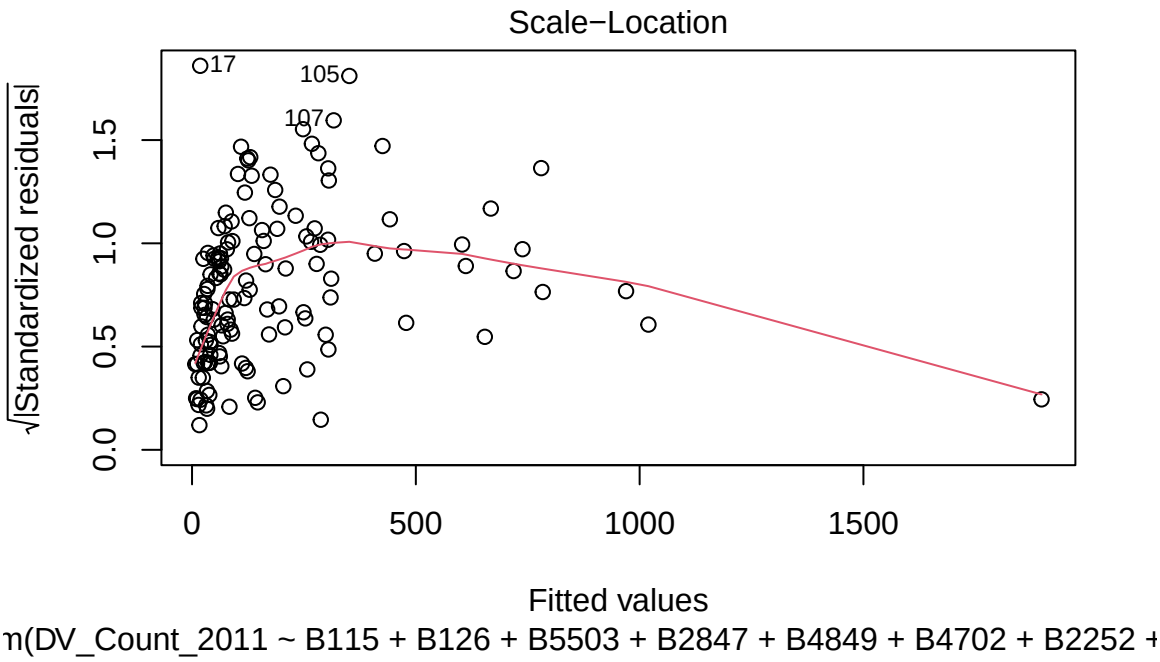
The model summary shows, the features added manually B115, B126, B5503 and B2847 has P-Value greater than 0.05, hance these are not a good predictor in the model. Therefore, we cannot reject H_0, i.e. There is no relationship between income, education and employment with DV. The features selected by Forward stepwise Feature selection algorithm are the best predictor to predict DV. The regression summary we can see these features have very low P-Value indicating the features having very strong relationship with the response variable. This fact is marked by *** in the summary. Also, the Multiple R-squared of the model returned 0.9831, indicating the predictors can explain 98.12% of variance from the model.

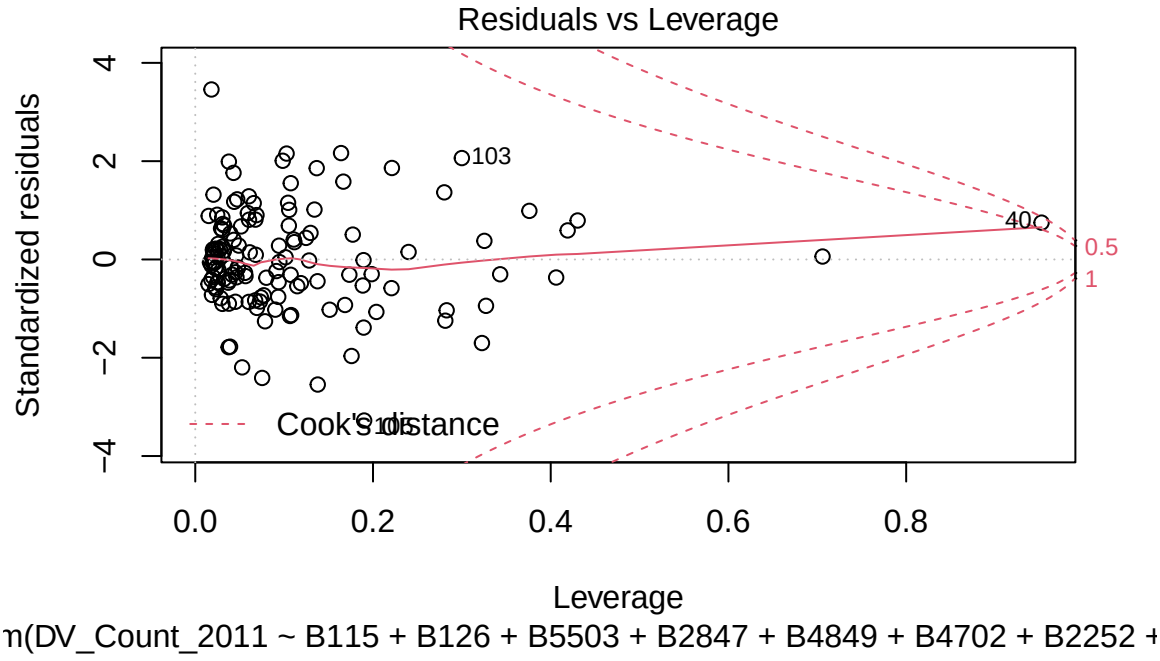


ɱ(DV_Count_2011 ~ B115 + B126 + B5503 + B2847 + B4849 + B4702 + B2252 +



ɱ(DV_Count_2011 ~ B115 + B126 + B5503 + B2847 + B4849 + B4702 + B2252 +

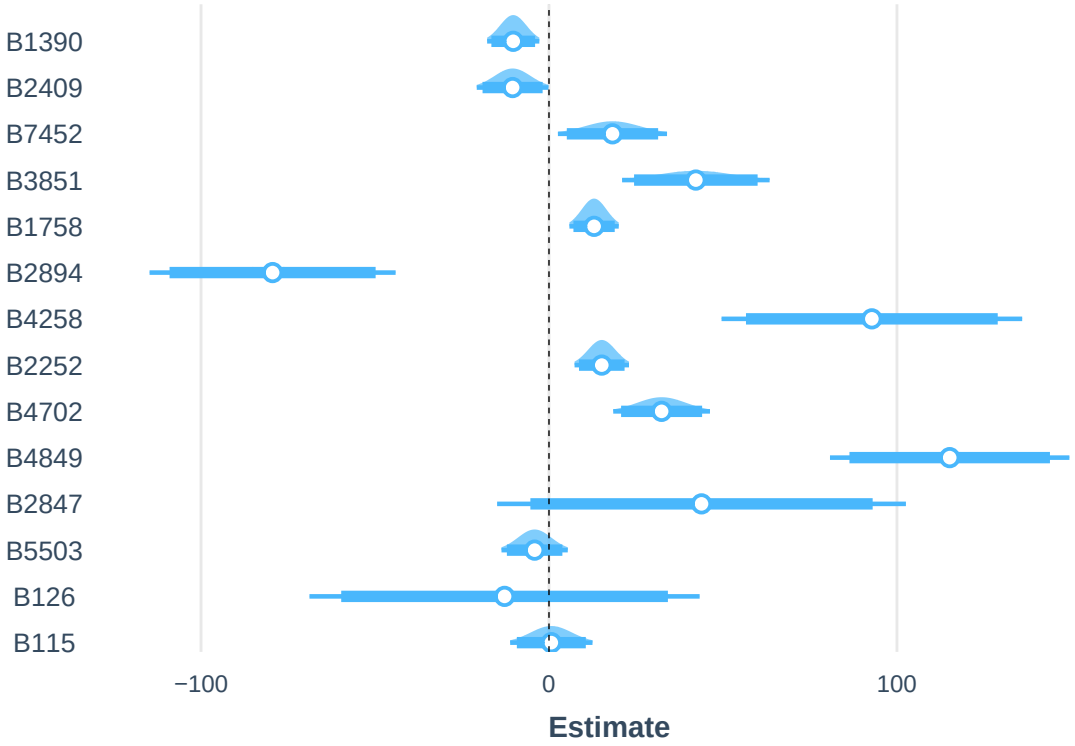

Scale−Location
17
105
107
√|Standardized residuals|
0.0
0.5
1.0
1.5
0
500
1000
1500
Fitted values
m(DV_Count_2011 ~ B115 + B126 + B5503 + B2847 + B4849 + B4702 + B2252 +


Residuals vs Leverage
Standardized residuals
4
2
0
−2
−4
103
40
0.5
1
Cook's distance
0.0
0.2
0.4
0.6
0.8
Leverage
m(DV_Count_2011 ~ B115 + B126 + B5503 + B2847 + B4849 + B4702 + B2252 +

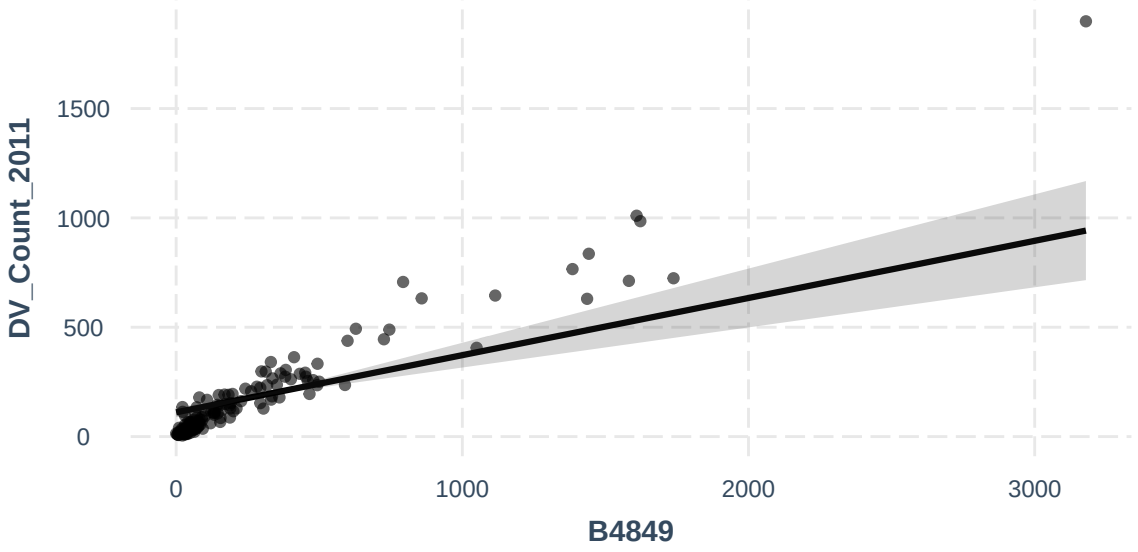
One of the best ways to present finding from a model is to generate plot_summs from jtools library. Using the plot, we can show the best predictor's coefficient estimations are normally distributed as presented in plot below.

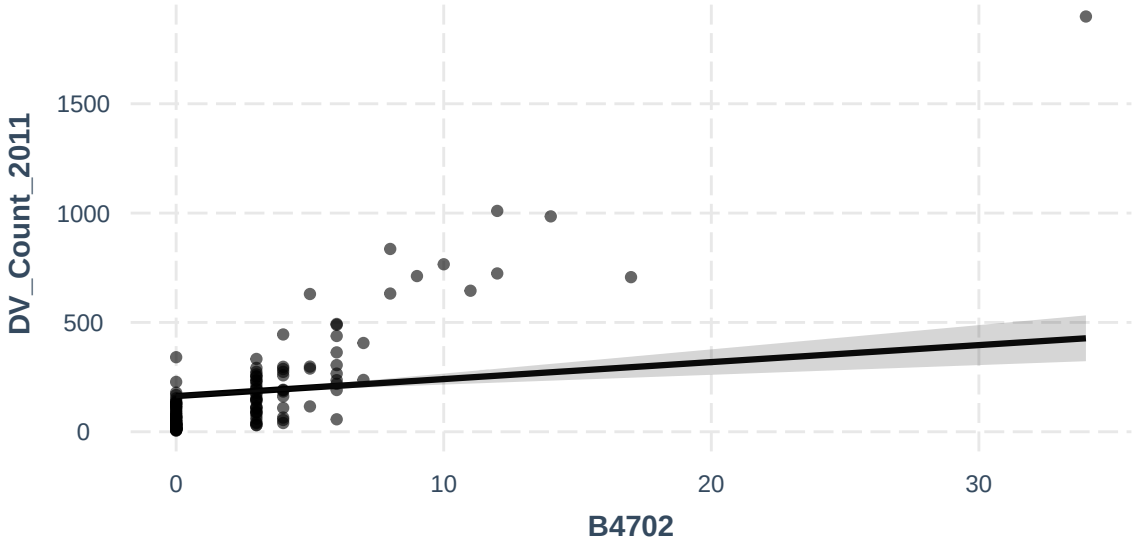
```
## Registered S3 methods overwritten by 'broom':
##   method            from
##   tidy.glht         jtools
##   tidy.summary.glht jtools
```

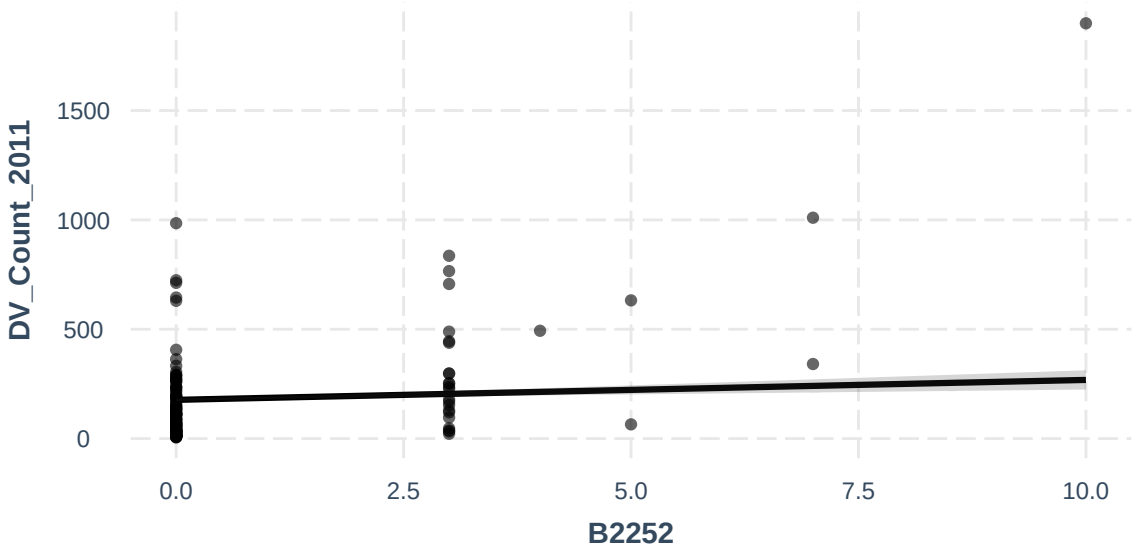
```
## Loading required namespace: broom.mixed
## Loading required namespace: broom.mixed
```

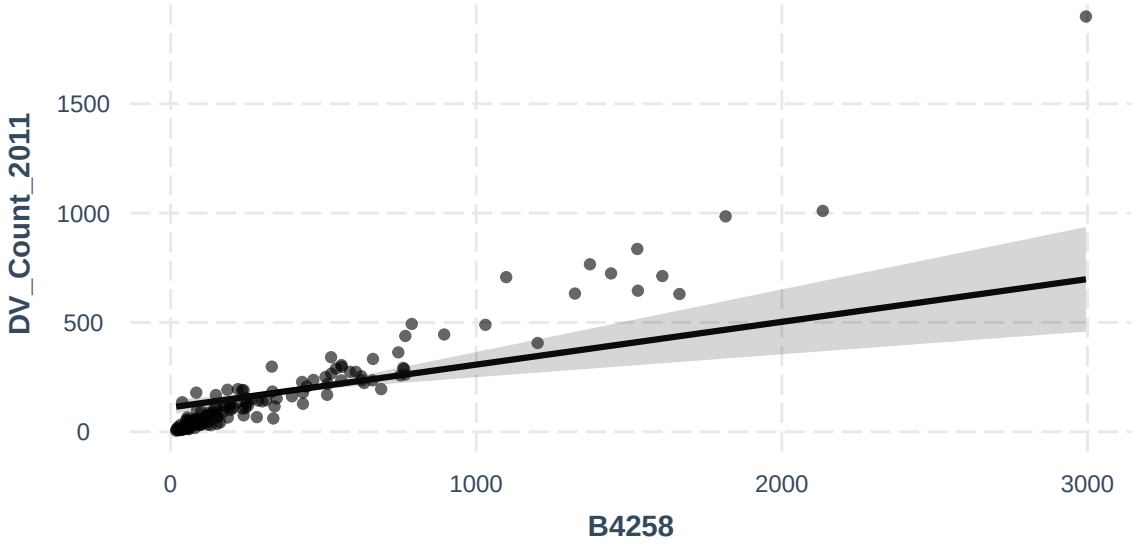


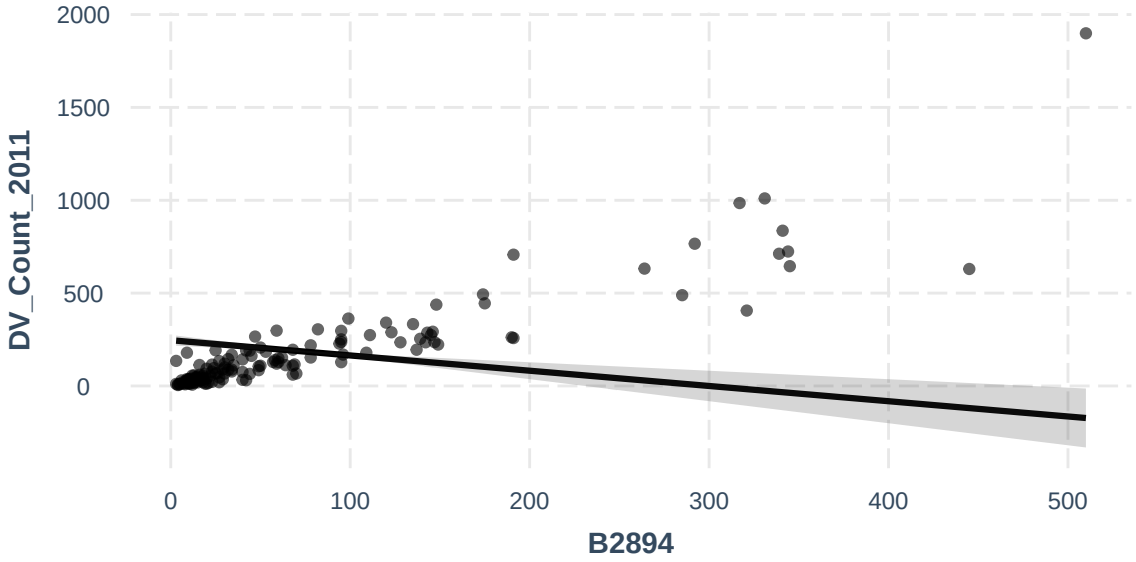
Another useful visualization is effect_plot to show the relationship between predictor and response variable. The predictors with low P-Value are visualized below. All these plot display very strong connection with the response variable.

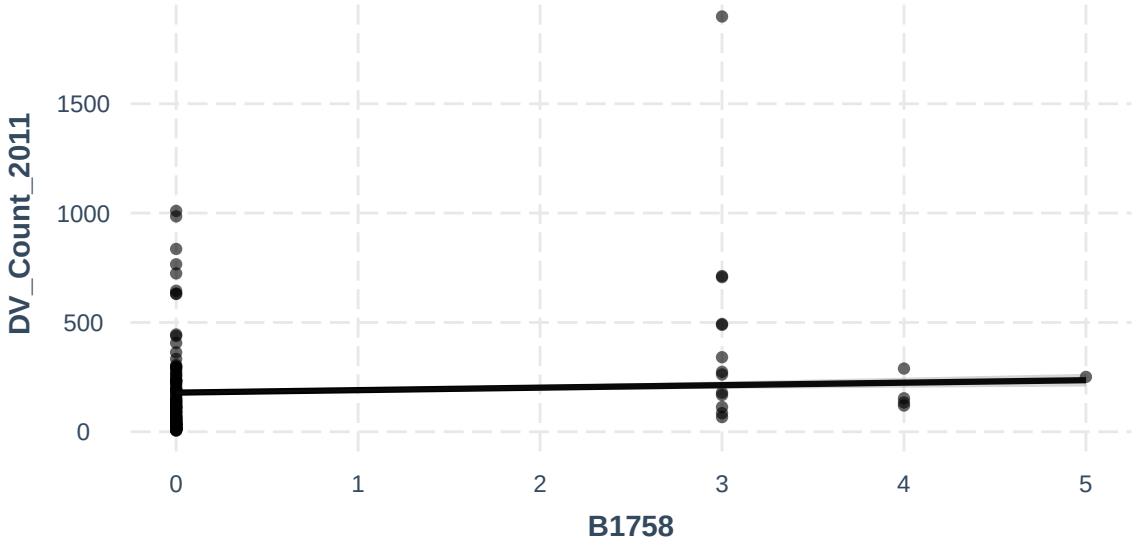

DV_Count_2011
1500
1000
500
0
0
1000
2000
3000
B4849

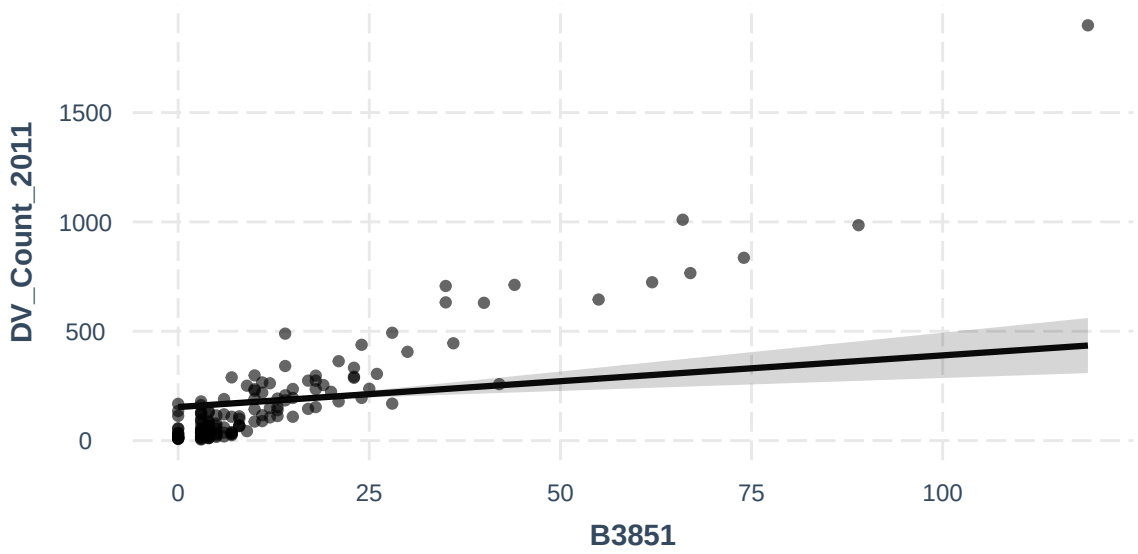

DV_Count_2011
1500
1000
500
0
0
10
20
30
B4702

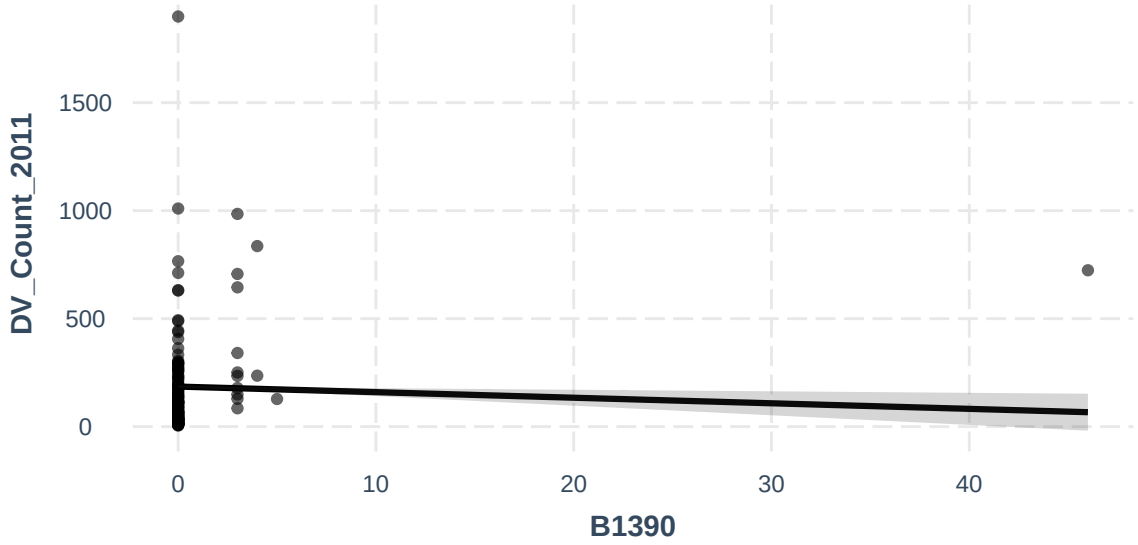

DV_Count_2011
1500
1000
500
0
0.0
2.5
5.0
7.5
10.0
B2252


DV_Count_2011
1500
1000
500
0
0
1000
2000
3000
B4258


DV_Count_2011
2000
1500
1000
500
0
0
100
200
300
400
500
B2894


DV_Count_2011
1500
1000
500
0
0
1
2
3
4
5
B1758





Conclusion

DV is a social problem in NSW. The data we have used in this project, clearly shows the level of crime does not have a downward trend. Policy maker and law enforcing agencies need to have more clear insight on the crime, so that they can come up with a policy to reduce this crime. This project can add a very good value in this regard. We have shown the trend of the crime and the area where the crime is happening the most. We have also identified some key predictors and build a regression model that can predict the crime very with very high level of confidence. The data we have worked on is not enough to build a complete model. We had only census data of 2011. In future we shall add more census and DV data and try to build a complete model that can predict DV with higher accuracy.