A Project on Crime

This project investigates whether the likelihood of someone reconvicting a crime can be predicted initially through the use of exploratory statistics and then using generalized linear models (GLMs). The data for this project is gathered from a job centre specializing in finding work for convicts

Julian Villar

Lancaster University
MATH333 Project

Contents

1 Introduction **1**

2 The Investigation **1**

3 Data Analysis **1**

3.1 Exploratory Analysis 1

3.2 Theoretical Analysis 3

3.3 Model Fitting 3

4 Conclusion **5**

1 Introduction

Modern day statistics often revolves around modeling real world problems to aid understanding of relations and prediction. This is most commonly used in fiance but we will focus on a social issue, crime.

Understanding crime is vital. Knowing what variables causes someone to be at a higher risk of committing crime not only helps us catch more criminals but also helps us in preventing the crimes from happening in the first place.

This report will focus particularly on what variables are important in predicting the likelihood of a convict reoffending.

2 The Investigation

This investigation is based on data gathered from a job centre that specialises in finding work for convicts, we are particularly interested in comparing what happened to the individuals in the two years following their visit to the centre.

The most important variable is obviously whether they reconvict `reconv` or not, but we will check the reconvinction rate compared to the number of previous convictions `precon` (taking an integer value), the convicts' age `age` (taking discrete values) and if they received a work placement `placed` (taking binary value where 1: placed, 0: not placed). For quantitative data like age and number of previous convictions it can be useful to have these in groups. Therefore for age we'll have three groups: young, middle-aged and old and for previous convictions they'll be grouped as low ($\texttt{precon} \leq 1$), medium ($2 \leq \texttt{precon} \leq 4$) and high ($\texttt{precon} \geq 5$).

Our data set consists of 347 convicts and we have no missing data points, thus methods for dealing with missing data will not be considered in this investigation.

3 Data Analysis

3.1 Exploratory Analysis

We will begin with exploratory data analysis. This is a vital step as drawing preliminary conclusions about what may be the biggest risk factors can cut down a lot of time as sometimes investigations have many more variables than this one.

It seems logical to first begin with just *how* many convicts have re-offended in a two year period. Using `sum(crime$reconv == 1)` we get 127 of the 347 convicts have reconvicted, that is 36.6% of convicts were found to have reconvicted after two years.

We'll begin the variable analysis with exploring whether work placement seems to have a significant impact on reconviction, we'll use relative risk to assess this. To do this we produce that following table.

	Placed	Not Placed	Total
Reconvicted	47	80	127
Not Reconvicted	112	108	220
Total	159	188	347

Table 1: Comparing Reconviction to Work Placement

Directly from the table we can see that 159 convicts were placed into work and 188 were not, so we can directly compare the proportion of convicts who reconvicted and **were** placed to those convicts who reconvicted and **were not** placed. So,

$$\frac{47}{159}\Big/\frac{80}{188} = \frac{47 \cdot 188}{159 \cdot 80} \approx 69.5\%.$$

This means that people placed in work reconvict at 69.5% of the rate at which convicts not placed in work reconvict, this is also called the Relative Risk.

Next it is important to understand the age distributions of our data and how that effects reconviction, we'll plot box plots to show this

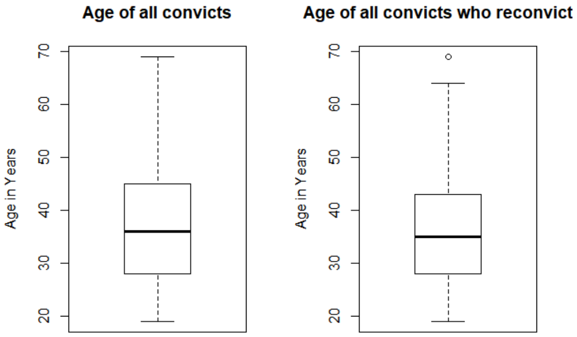

Figure 1: Box Plots showing age distributions of all convicts compared to reconvicts.

The distribution of age in these plots seem quite similar. The reconvicts have a slightly smaller interquartile range which would suggest that most of the convicts lie between the ages of about 28-42, the reconvicts also have a slightly lower median. Overall not much conclusion can be drawn from this, but due to the potential dependency on other variables that age may have it's unreliable. For example we might see that older people may have a more difficult time getting jobs.

Finally we will look at the effect of previous convictions on reconviction. Graphically it is much easier to visualise `GPRECON` as we have gaps between values of `precon`. The following histograms show the distribution of grouped previous convictions of all convicts compared to the distribution of grouped previous convictions of those who went on to reconvict

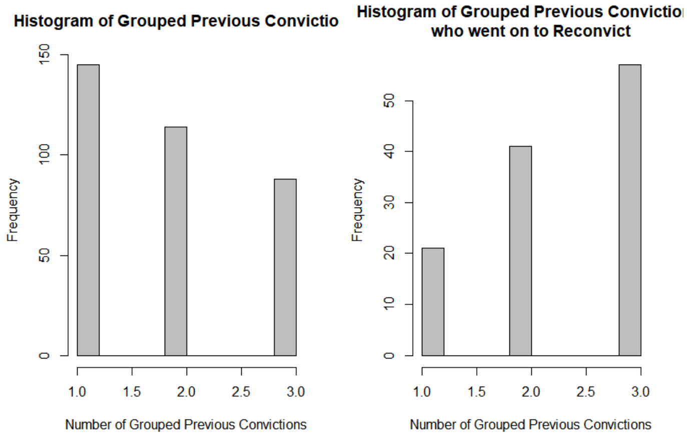

Figure 2: Histograms comparing Previous convictions of all convicts compared to reconvicts.

These two graphs are very different in what they show, initially from all of the convicts those who have committed ≥ 5 crimes seem to be in the minority but from the reconvicts they are very much the majority. This suggests that number of previous convictions has a large reason as to why someone would reconvict.

3.2 Theoretical Analysis

For further analysis of data we need a more rigorous approach. Typically when you think of the structure of a model you have an output or *response* which corresponds to an input or *explanatory variable*. These models are known as Generalized Linear Models, or GLMs. We assume that the observations for the model are taken on a one-dimensional response variable Y_i, which is indexed by i from $i = 1, \dots, n$. Whereas the explanatory variables depend on some p that is strictly less than n or $x_{1,i}, \dots, x_{p,i}$. We call the responses dependent on i the realisations, these are assumed to be observed independently.

We condition that Y is part of the exponential family and we require a mean μ and a known constant scale parameter ϕ.

We have a linear predictor which is a linear function dependent on the explanatory variables

$$\eta = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p. \tag{1}$$

Here our β values are coefficients for the response variable. The linear predictor and the mean of the response variable are related by the link function, that is that a function g of the mean μ equals the linear predictor η or $g(\mu) = \eta$.

3.3 Model Fitting

Typically with models you have a high number of variables and checking every single combination of variables grows exponentially more difficult as the models become more complex, therefore we must look to more general methods. We will use the forward selection procedure, whereby we assess which explanatory variable to add next by adding the one that reduces the residual deviance to a significant level the most. After this we trial the interactions to assess whether they are significant.

We begin with the null model, our family is binomial and our link function is logit(μ) that is in `R: Null = glm(reconv ~ 1, family = binomial, data=crime)`. This outputs a Null deviance of 455.81. The further models will be named in the form of what they are dependent on e.g dependency on age is `DepAge`.

The choices of possible models are as follows,

```
DepPrecon = glm(reconv ~ 1 + precon, family = binomial, data=crime)
DepGPRECON = glm(reconv ~ 1 + GPRECON, family = binomial, data=crime)
DepAge = glm(reconv ~ 1 + age, family = binomial, data=crime)
DepGAGE = glm(reconv ~ 1 + GPRECON, family = binomial, data=crime)
DepPlaced = glm(reconv ~ 1 + placed, family = binomial, data=crime).
```

Finding the `summary` of these models produces the following residual deviances:

Model	Deviance	df.	Test Stat	df	Accept/Reject H_0
Null Model	455.8	346	-	-	-
`precon`	402.4	345	53.4	1	Reject
`GPRECON`	402.7	345	53.1	1	Reject
`age`	447.7	345	8.1	1	Reject
`GAGE`	452	345	3.8	1	Accept
`placed`	449.5	345	4.3	1	Reject

Table 2: Residual Deviances of first explanatory variable.

Straight away we see that previous convictions has a much bigger effect than every other variable, using χ_1^2 we get a significance level of 3.84. In this context that means the residual deviance **difference** must be at least 3.84 to be significant enough to reject H_0. For further model fitting we will void the `GPRECON` variable as it is linearly dependent to `precon` and it wouldn't make sense to have both in a model. Continuing model fitting we find,

Model	Deviance	df.	Test Stat	df	Accept/Reject H_0
`precon`	402.4	345	-	-	-
`precon + age`	389.6	344	12.8	1	Reject
`precon + GAGE`	395.7	344	6.7	1	Reject
`precon + placed`	401.1	344	1.3	1	Accept

Table 3: Residual Deviances of second explanatory variable.

Here the biggest reducer of deviance is the `precon + age` model. As before we discontinue `GAGE` from future model fitting, therefore our final iteration of new explanatory variables is

Model	Deviance	df.	Test Stat	df	Accept/Reject H_0
`precon + age`	389.6	344	-	-	-
`precon + age + placed`	388.6	343	1	1	Accept

Table 4: Residual Deviances of third explanatory variable.

Since H_0 is accepted we won't include `placed` in any model fitting at all since it adds unnecessary complexity. Finally we have to check the interactions of this model, that is the interaction term `precon:age`

Model	Deviance	df.	Test Stat	df	Accept/Reject H_0
`precon + age`	389.6	344	-	-	-
`precon + age + precon:age`	389.2	343	1	1	Accept

Table 5: Residual Deviances of all interaction terms

The interaction term is not significant so our final model in predicting crime is dependent on only `precon` and `age`. Since our link function is logit(μ) we know our linear predictor is

$$\text{logit}(\mu) = \eta = 0.09604 + 0.27609x_1 - 0.04394x_2. \tag{2}$$

We use $\text{logit}^{-1}(\mu)$ or logistic(μ) to find the probability for a specific convict. For example a 25 year old with 8 previous convictions has the following probability associated with reconviction

$$\text{logit}^{-1}(\mu) = \text{logistic}(\mu) = \frac{1}{1 + \exp(-(0.09604 + 0.27609 * 8 - 0.04394 * 25))} = 0.76964.$$

4 Conclusion

To conclude, we have found that the best model to fit for prediction of crime is

```
glm = (reconv ~ 1 + precon + age, family = binomial, data = crime).
```

Throughout this investigation we have seen that whether someone is placed into work or not has no significant effect on reconviction, we have found that age has a significant effect (albeit small) but the overwhelmingly most important variable is the number of previous convictions. Therefore the high-risk group are people with numerous previous convictions.

However, we could potentially improve this investigation massively by tracking other variables. For example, we could see if violet crime offenders are more likely to reconvict compared to drug offenders. We could also look into how educated they are, how poor they may be and whether they have direct family bonds with criminals.

The summary of this report is not particularly enlightening, we found that people who have gone on to reconvict several times in the past continue to do so however with more data and more variables we may be able to target a more specific high risk group in a different investigation.