psyc30815 2020/21 Assignment

Jon Doe

January 11th, 2020

Question 2: Logistic Regression

This analysis concerns factors that predict an arrest outcome for marijuana possession.

Data-set

The data set I have chosen for this analysis is based on police officers' treatment of individuals for the possession of small quantities of marijuana in Toronto, Canada. The data are part of a larger data set featured in a series of articles in the Toronto Star newspaper. The data featured was collected between 1997 and 2002. This data is available as the `Arrests` data-set in the `carData` R package.

I decided that the most interesting and compatible *outcome* variable was the one called **released.**

released Whether or not the arrestee was released with a summons; a binary variable with levels: No; Yes.

The following *predictor* variables from the data set were deemed to be worth analysing.

colour The arrestee's ethnicity; a factor with levels: Black; White.

year 1997 through 2002; a numeric vector.

age in years; a numeric vector.

sex a factor with levels: Female; Male.

employed a factor with levels: No; Yes.

citizen a factor with levels: No; Yes.

checks Number of police data bases (of previous arrests, previous convictions, parole status, etc. – 6 in all) on which the arrestee's name appeared; a numeric vector

Objective of analysis

The data set is interesting partly because of the Black Lives Matter movement in the US. Proponents of the movement argue that African Americans are targeted both unfairly and disproportionatly by the police. Whilst this data set is from Canada, it would be interesting to see if the kind of disparities the BLM movement argues exist, occur in Canada on the issue of drug possession. There have also been discussions around African Americans being unfairly punished for drugs crimes compared to caucasian people in the last century; the Netflix documentary '13' covers this topic in detail. It is also interesting because marijuana is now being legalised increasingly across the world. Identifying which factors are most likely to get you charged for possesing a drug which is viewed so differently across nations, and even states, is both topical and interesting.

The focus of the analysis will therefore be on whether, all else being equal, the colour of someone's skin affects the probability that they are released on a summons for the possession of a small quantity of marijuana. However, there are various other factors that will be interesting to examine in this exploratory context, therefore they will be included as covariates.

Preparation of the data and assumptions

The data is already compatible with the assumptions for a logistic regression. The outcome variable is binary and in this case, *yes* or *no* regarding being released on a summons.

The observations are independent of each other and not done by repeated measures.

The is no reason to suspect any multicolinearity between the independent variables.

There are sufficient observations of the data, 5226.

Analysis

```
url <- 'http://vincentarelbundock.github.io/Rdatasets/csv/carData/Arrests.csv'
weed_df <- read_csv(url) %>% mutate(released = released == 'Yes')

logm <- glm(released ~ colour + checks + year + age + sex + employed + citizen,
            data = weed_df,
            family = binomial)


logmnull <- glm(released ~ 1, data = weed_df, family = binomial)

logm2 <- glm(released ~ colour + checks + employed + citizen,
             data = weed_df,
             family = binomial)

logm3 <- glm(released ~ colour + checks * citizen + employed,
          data = weed_df,
          family = binomial)
```

```
coefs <- logm3$coefficients
```

Formally, our assumed binary logistic model can be written as follows: For $i \in \{1 \ldots N\}$,

$$\log\left(\frac{\theta_i}{1-\theta_i}\right) = \beta_0 + \sum_{k=1}^{K=7} \beta_k x_{ki}$$

Where θ is the probability, a conditional mean of y. $x_1i...x_ki$ are the values of the $K = 7$ predictor variables in observation i. The to-be-inferred parameters of the model are the intercept term β_0, the coefficients of each of the $K = 7$ predictor variables, and the deviance of the model. The deviance explains how good the model-fit is, and it uses the maximum likelihood estimation. It is denoted below:

$$Deviance = -2 \log L(\hat{\alpha}, \hat{\beta}|D)$$

where $\hat{\alpha}, \hat{\beta}$ are the maximum likelihood estimates. D is a set of observed data whose probability distribution is parameterized by θ.

To calcuate a probability from the to-be-inferred coefficients, the following formula can be used. z is the logit function, however since the desired probability would be unknown, we will use the linear function in the logit equation as z, as it is equivalent. This equation is called the inverse logit function.

$$\theta_i = \frac{1}{1+e^{-z_i}}, \quad z_i = \beta_0 + \sum_{k=1}^{K=7} \beta_k x_{ki}.$$

The overall model fit of this binary logistic model using the $K = 7$ predictors on the *released* outcome variable was deviance $D = 4,299$. The null hypothesis was then tested by comparing this model to the null model. The null model's deviance was $D = 4,776$. The p value corresponding to the global null hypothesis that all predictor variables are simultaneously of no predictive value is highly signifcant, at $P < 0.001$. This indicates that the model can predict, to some extent, the outcome variable *released*.

The maximum likelihood estimates, standard errors, z-statistic, and p-values for the coefficients are provided in the following table:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	9.372	56.72	0.1652	0.8688
colourWhite	0.3891	0.08566	4.542	5.565e-06
checks	-0.3641	0.02598	-14.01	1.303e-44
year	-0.004218	0.02838	-0.1487	0.8818
age	0.002236	0.004631	0.4827	0.6293
sexMale	0.007317	0.1502	0.04872	0.9611
employedYes	0.7573	0.08474	8.937	3.988e-19
citizenYes	0.5765	0.1042	5.53	3.196e-08

It is clear that four of the seven predictor variables are highly significant (year, age and sex were not). If all the other predictor variables were held constant, *colour*, *employed*, *citizen* and *checks* would all reliably predict a change in the outcome variables, *released*. This means the coefficients for the significant predictor variables would change the logit odds of the outcome variable *released* being TRUE, which in this case relates to the chance that an individual was released on a summons for marijuana possession. Year, age and sex appear, due to their high p-values, to have no predictive value. This can be checked by dropping them, running a new model, then comparing the deviance to the original model. The new model, $K = 4$ was run and it was found to have a deviance of $D = 4,299$, the same as model $K = 7$. A model comparison was then performed between these two models and the p-value was 0.97, indicating that there is no significant different in model fit between the $K = 4$ model and the $K = 7$ model.

The model was then tested for interaction effects. A two-way interaction between *checks* and *citizen* was found, so this was put into a new model and this was run. The $K = 4$ interaction model had a deviance of $D = 4,287$, which is 12 lower than the non-interaction $K = 4$ model. These two models were compared using an ANOVA and the difference between them was significant $p < 0.001$. It can therefore be concluded that the $K = 4$ model, with an interaction between *checks* and *citizen* has the best model fit.

The maximum likelihood estimates, standard errors, z-statistic, and p-values for the coefficients in the $K = 4$ interaction model are provided in the following table:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.6685	0.1546	4.324	1.53e-05
colourWhite	0.3902	0.08496	4.593	4.363e-06
checks	-0.2035	0.05177	-3.932	8.427e-05
citizenYes	1.018	0.1586	6.417	1.394e-10
employedYes	0.7479	0.0839	8.915	4.874e-19
checks:citizenYes	-0.2069	0.05903	-3.506	0.0004557

The variables *colourWhite*, *citizenYes* and *employedYes* all have positive slopes. This means that all else being held constant, if the arrestee has either white skin, Canadian citizenship, or is currently employed, the odds ratio for them being released on a summons increases. In contrast to this, the variable *checks* has a negative slope. This means that, all else held constant, a unit increase in the variable *checks* (number of police databases they appear in) will cause the odds ratio for them being released to decrease.

The interaction effect also has a negative slope. This means that for a unit change in *checks* when *citizen* is 'No,' the odds ratio change is $exp(-0.2035)$, which is 0.82, but when citizen is 'Yes,' the odds ratio change is $exp(0.2035 - 0.2069)$, which is 0.66.

It is important to note that all the inferred coefficients presently indicate the change in logit odds to the outcome variable. The logit odds can be difficult to visualise mentally however, so I have described changes in terms of odds ratio change. This is acheived by doing the exponential function of the logit odds. Whilst the logit odds do not have a linear relationship with the odds ratio, the magnitude and direction of the logit odds coefficient (positive or negative), both hold true when it is converted to an odds ratio. If the coefficient is negative, the odds ratio will decrease in size, if it is positive, the odds ratio will increase in size. The logit odds however, do vary as a linear function of predictor variables.

Predictions

Using this $K = 4$ interaction model, it is possible to make predictions about the data. For this prediction, we will see how *colour* and *checks* affect one's chance of being *released.* We will also visualise *citizen* and *checks*, so the interaction can be observed.

Below is a table of the probabilities of being *released* for different levels of *checks* and *colour* whilst citizen and employed are kept constant.

Below, **Figure 3** shows two graphs. The **left-hand** graph shows two negative slopes, which mean that as the number of *checks* increase, the probability of being *released* decreases. However, the intercept value is different depending on whether *colour* is Black or White. The graph shows that being White, for all levels of *checks*, increases the probability that the person will be *released.* The **right-hand** graph shows two negative slopes and an interaction effect between *citizen* and *checks.* Whilst being a *citizen* significantly increases your chances of being *released* when *checks* equals 0, for each additional *check* you have, the probability of being *released* decreases at a faster rate than if you were not a citizen. This actually means that for *checks* = 5 or 6, being a citizen make you less likely to be released, than if you were not.

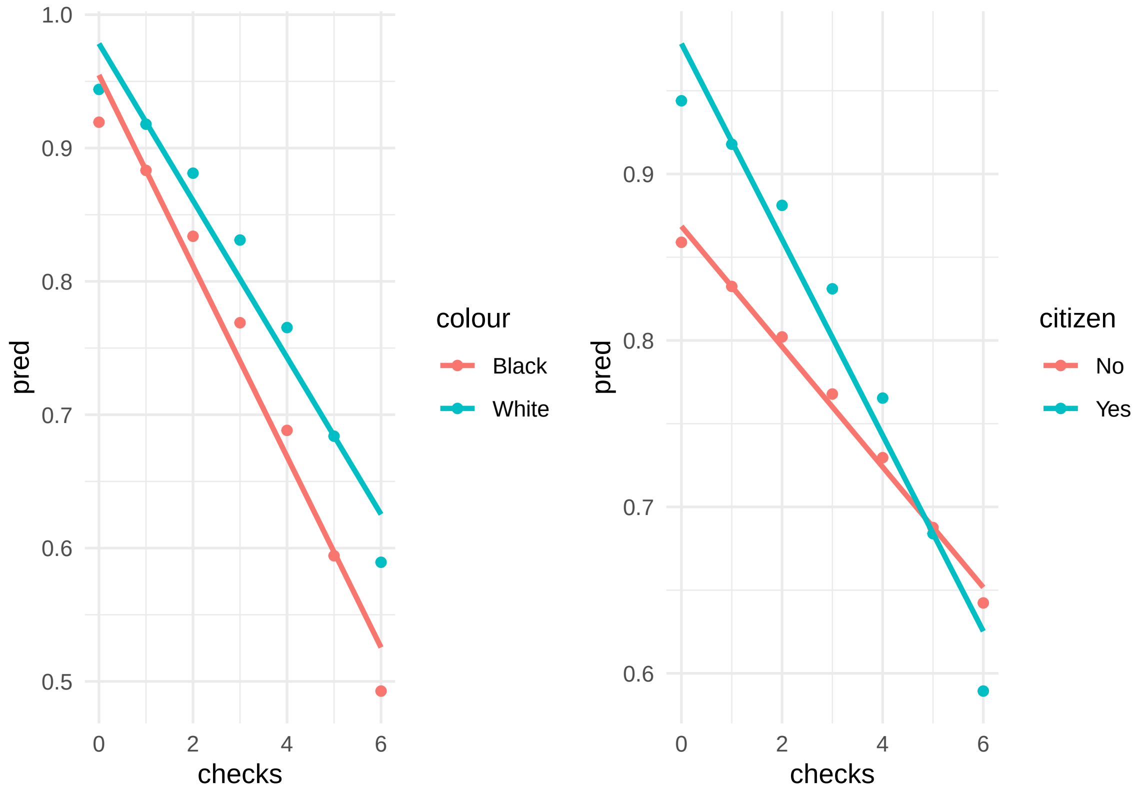

Figure 1: Predicting the probability of *release* from *checks* and *colour*

checks	colour	employed	citizen	pred
0	White	Yes	Yes	0.9439766
0	Black	Yes	Yes	0.9193903
1	White	Yes	Yes	0.9178764
1	Black	Yes	Yes	0.8832526
2	White	Yes	Yes	0.8811478
2	Black	Yes	Yes	0.8338421
3	White	Yes	Yes	0.8310168
3	Black	Yes	Yes	0.7689891
4	White	Yes	Yes	0.7653715
4	Black	Yes	Yes	0.6882861
5	White	Yes	Yes	0.6839241
5	Black	Yes	Yes	0.5942651
6	White	Yes	Yes	0.5893723
6	Black	Yes	Yes	0.4927835

What factors predict release from a marijuana arrest?

This is a real data set and as such the results have real-world implications. This analysis has shown that having black skin makes you slightly less likely to be released for a possessing a small quantity of marijuana, than someone who has white skin. It also shows that, as you might expect, the more time your name appears on police databases, the less likely it is you will be released. Furthermore, being a citizen and being in employment increases the likelihood that that you will be released.

All these conclusions should be considered in the content of this data: Toronto - Canada, 1997 - 2002. Warren, Chiricos, and Bales (2012) confirm the sentiment of the findings here, as they found that in Florida, USA, black and hispanic rates of incarceration are 6 times higher than that of white offenders. They also found that women were much less likely to be incarcerated for offending than men. This is a slight contrast to the present analysis, where sex did not influence being released on a summons. However, it's important to realise that being incarcerated is significantly different to being charged with possension of a small quantity of marijuana.

References

Warren, Patricia, Ted Chiricos, and William Bales. 2012. "The Imprisonment Penalty for Young Black and Hispanic Males: A Crime-Specific Analysis." *Journal of Research in Crime and Delinquency* 49 (1): 56–80.