Sports Rorts

Jeremy Forbes

19/02/2020

## Uncovering government bias with advanced statistical modelling.

Pork barrelling - a metaphor for government spending solely or primarily to bring money to a representative’s district.

In Australia, in not unusual to see pork barrelling accusations hurled between the two major parties. In recent weeks, the Liberal party has been under intense scrutiny for its allocation of grants to sporting organisations under a scheme that took place leading up to last year’s federal election. Dubbed the “sports rorts” affair, the Liberal party is accused of approving grants on the basis of an underlying strategy that aimed to boost support in targeted electorates.

As a data scientist, I see this accusation as a question that could be readily explored using publicly available data and a pragmatic approach to statistical analysis. My theory was that if electorates were in fact preferenced based on their previous voting behaviour, then the Liberal party’s strategy would be revealed by fitting an appropriate statistical model.

Now before I go any further, here’s some background incase you haven’t read the news…

#### Background

The Australian Liberal party has been in power since the 2013 federal election. In the lead up to last year’s election, the Liberal party announced a community sport infrastructure program, offering grants of up to $500,000 for local sporting organisations. This program was marketed as an intitiative to boost sporting participation across the country. Organisations had to submit an application to be independently assessed by Sport Australia, from which they were given a score. These scores were to be used by Sports Minister in deciding which grants would be approved.

This year an audit on the funding scheme was conducted by the Australian National Audit Office (ANAO). Their report implied that the government, led by the Liberal party, gave preference to grant applications from ‘marginal’ and ‘targeted’ electorates, rather than following the advice from Sport Australia. Essentially, the Liberal party are accused of adopting a strategic plan to win support in specific electorates (seats) by approving more grants in those areas, to order to help them secure victory in the 2019 election.

#### Back to the action

My aim is to explore whether there is evidence to suggest that the Liberal party did target electorates on the basis of their previous voting behaviour. Essentially, is there an underlying strategy? And if so, what voting pattern was targeted?

To go ahead with this analysis, I will need to combine three key data sources, all of which are publicly available: (1) Information about each grant that was approved (location and $ amount), (2) Election results for each electorate from the previous election, and (3) Census data to control for other socio-demographic factors that might be at play.

With this information, I was confident I would be able to find evidence to support or oppose the accusations. Here’s how I went about each step.

(All steps of data collection, wrangling and analysis were conducted using R.)

#### Obtaining the data

Sports Australia published a list of the 684 approved grants in a table on their website (<https://www.sportaus.gov.au/grants_and_funding/community_sport_infrastructure_grant_program/successful_grant_recipient_list>), which I scraped using *rvest*. I then look this list and fed it to a Google Maps API to get the lat-long coordinates of each grant (using *ggmap* from <https://github.com/dkahle/ggmap>). This worked for almost all of the grants, so I did some quick googling to see find where the remainder of these organisations were located.

To allocate each sporting organisation to its corresponding seat, I would need a map of electorates from the 2016 federal election. Luckily, this is readily available in the eechidna package (which I maintain), along with the electoral voting data. I used some nifty functions from sp and rgeos to loop through the electorates and complete this allocation.

Next I aggregated the grants, so that the resultant dataset would hold the total grant $ amount for each of the 150 federal electorates. This was then joined with the two-party preferred vote from the 2016 election, along with a collection of socio-demographic variables from the 2016 Census - also obtained from eechidna.

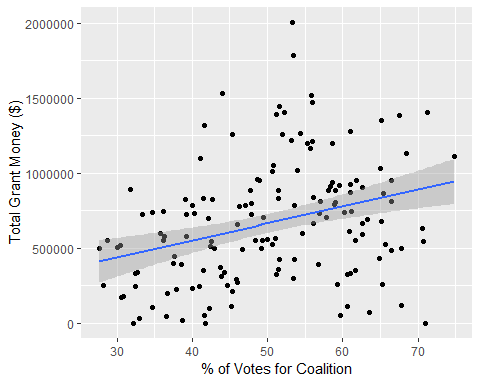
Here’s a snapshot of the resultant dataset.

## # A tibble: 6 x 10  
## DivisionNm Population MedianAge MedianPersonalI~ HighSchool Unemployed  
## <chr> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 ADELAIDE 163442 36 463. 66.3 8.1  
## 2 ASTON 136018 39 455. 58.5 5.6  
## 3 BALLARAT 154483 40 406. 45.8 6.5  
## 4 BANKS 155806 38 441. 62.0 6.2  
## 5 BARKER 149502 44 390. 34.1 5.8  
## 6 BARTON 172850 35 436. 65.7 6.6  
## # ... with 4 more variables: Owned <dbl>, Swing <dbl>, Amount <dbl>,  
## # LNP\_Percent <dbl>

Now time to get stuck into the analysis.

#### A Statistical Modeling Approach

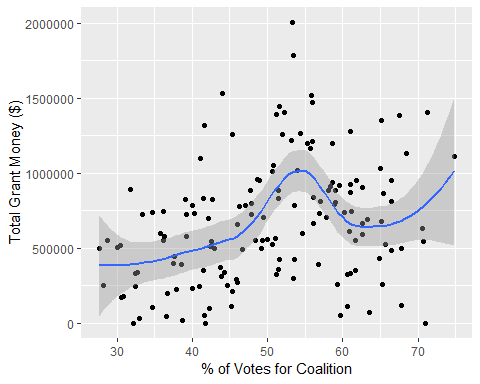
What I’m really looking to uncover is the relationship between the percentage of votes in favour of the Liberal party for each electorate in the 2016 election, and the amount of grant money allocated to that electorate. First, let’s see what a scatterplot looks like with a linear regression line fitted.



There’s some evidence here to suggest that electorates with stronger support for the Coalition were more likely to receive sports grant money. But really, its likely that if the Coalition did implement a targeted strategy in allocating grants, the relationship between these two variables would not be linear. A strategy that gives the most money to seats that are the safest wouldn’t be a very good one - because its highly unlikely that these seats would be vulnerable to changing hands at the 2019 election.

So instead of fitting a linear regression line, let’s plot a really flexible regression curve. For this I have chosen to use the non-parametric loess curve, which gives higher weighting to local data points to fit a smooth curve. This allows us to get a better picture of the relationship between the 2016 vote and grant amount, rather than imposing a linear restriction.

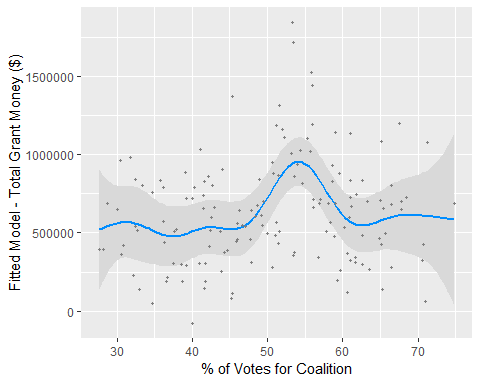
## `geom\_smooth()` using method = 'loess' and formula 'y ~ x'



Wow! The loess curve is revealing a pretty significant relationship, which was not obvious from inspecting the initial scatterplot. There is a clear spike in the amount of grant money given to electorates that were “marginal” Coalition seats in the 2016 election. “Marginal” Coalition seats are those that were won by the Coalition, but weren’t won by a huge margin. I think this plot itself is pretty convincing evidence of a deliberate strategy in allocating the grants.

But to take this another step further, I decided to fit a GAM (Generalized Additive Model) to model the grant amount as a function of the 2016 vote, Census characteristics of the electorates and the swing vote (the percentage point change in the vote from 2013 to 2016 elections for that seat). Each of these variables is included in the model as a smooth term, which are penalized regression splines - a type of non-parametric regression, similar to loess, that allows for patterns in the data to reveal themselves.

By controlling for these other variables, the effect of the vote itself can be identified. It is possible that once these other effects are accounted for, we will no longer see a relationship between the vote and grant amounts. Let’s go ahead and see…



It can be clearly seen that after accounting for these other variables, there is still a huge spike in grant amount for marginal Coalition seats! Having a look at the p-values from the GAM regression output, there is significant evidence that the 2016 vote influenced the allocation of grant money across electorates (p = 0.002). The resultant model itself explains 46.3% of the deviance - which is a good fit considering that the official aim of the grant scheme was set out to increase sport engagement Australia-wide, and the applications were scored by an independent body.

I also fit a model omitting the voting data (both two-party preferred and swing vote), and compared the two models with a Chi-square test. The model including the election data was a significantly better fit (p = 0.000005). On this basis, the evidence of a strategic bias is quite compelling.

To check that my findings are robust to outliers, I re-estimated the model after removing the two electorates with outstandingly large grant amounts ($1.8M and $2.0M - the electorates of Boothby and Dawson), and found that the same effects were observed. Additionally, the model assumptions seem to be valid, as the residuals are approximately normal (observed using a quantile-quantile plot).

#### In conclusion

My statistical modelling approach provides empirical evidence that the Liberal party did in fact allocate sports grants on the basis of an underlying strategy to beef up support in the lead up to the 2019 election. In a classic case of pork barrelling, marginal seats that were won by the Liberal party received significantly higher grant allocations, in what appears to be an attempt to secure these electorates for the 2019 election.

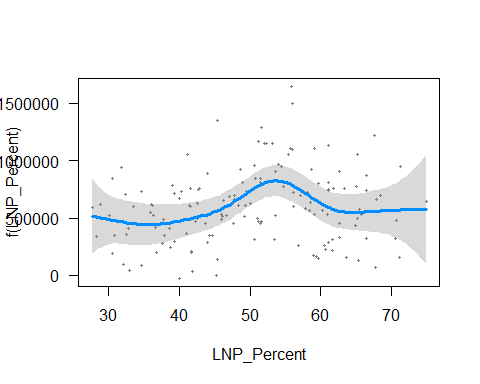
Perhaps equally importantly, this analysis demonstrates how powerful insights can be found using publicly available data.

All of the data collection, wrangling and analysis was conducted using R. For anyone interested in exploring Australian election and Census data, I highly recommend checking out the eechidna package (of which I am an author and maintainer). This package makes it easy to access Australian election and Census data from 2001-2019 (including maps), available on CRAN.

## REMOVE BELOW

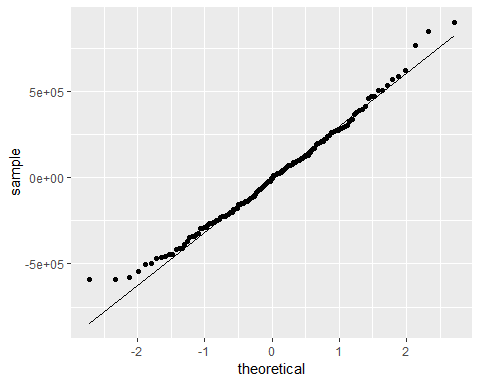
#### Checking that my findings are robust

Looking at the distribution of grant amounts across electorates with a simple boxplot reveals that there are two outliers with grant amounts of $1.8M and $2.0M - the electorates of Boothby and Dawson. I refit the model after removing those two electorates from the dataset, and still find that the effect of the 2016 vote is significant (p = 0.008), and the spike in grant amount still remains.



Similarly, when removing the seats with high Cook’s Distance and refitting the model, the same effect is observed.

There also don’t seem to be any material issues with the residuals as they appear approximately normal in a quantile-quantile plot.



Check out my #medium debut on my analysis of the “Sports Rots” scandal - a statistical modelling to see how sporting grant money was allocated on the basis of previous election results, all done using the #eechidna #Rstats package and a Google Maps API.