Sports Rorts

Jeremy Forbes

19/02/2020

## Sports Rorts - Uncovering Government Bias with Statistical Modelling

You can run, but you can’t hide (from statistics). The Australian Liberal party is finding this out the hard way.

In recent weeks, the Liberal party has been accused of using $100M of sporting grants to gain votes in the lead up to the 2019 election. As a data scientist, I believe there’s only one way to find out if this is true - by crunching the numbers. And so, I set out to see what can be discovered with publicly available data relating to the “Sports Rorts” scandal.

My thinking is that if the Liberal party allocated grants to electorates on the basis of their voting behaviour, then this strategy would be revealed with an appropriate statistical model.

Before I go any further, here’s some background in case you haven’t been reading the news…

#### Background

The Australian Liberal party has been in government since the 2013 federal election. In the lead up to last year’s election, the government announced a community sport infrastructure program, offering grants of up to $500,000 for local sporting organisations. Marketed as an initiative to boost sporting participation across the country, organisations had to submit an application to be independently assessed and scored by Sport Australia. These scores were supposed to be used by the Sports Minister in deciding which grants would be approved.

This year Australian National Audit Office conducted an audit and their report implied (but stopped short of asserting) that 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 supplying more grants in targeted areas in order to win votes for the 2019 election.

#### Back to the action

The ultimate aim of my analysis is to see whether there is evidence that grants were allocated on the basis of electorates’ voting behaviour from the previous election. Essentially, is there an underlying strategy? And if so, what voting pattern was targeted by the Liberals?

To go ahead with this analysis, I 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 (2016), and (3) Census data to control for other socio-demographic factors that might be at play.

With this information, I’ll be able to produce a dataset containing the total grant money allocated to each electorate, along with the corresponding voting results from the previous election and other controls from the census.

(Note that all steps of data collection, wrangling and analysis are conducted using R.)

#### Obtaining the data

Sports Australia has 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>), so I scraped this using *rvest*. Next, this list is fed to a Google Maps API in order to get the lat-long coordinates of each grant (using *ggmap* from <https://github.com/dkahle/ggmap>). This works for almost all of the grants. Some quick googling helps find where the remainder of these organisations are located.

To allocate each sporting organisation to its corresponding electorate, I need a map of electorates from the 2016 federal election. Conveniently, this is available in the eechidna package, along with the electoral voting data. Using some nifty functions from sp and rgeos to loop through the electorates, each grant is successfully allocated.

The grants are now aggregated so that the resultant dataset holds the total grant $ amount for each of the 150 federal electorates. This is then joined with the two-party preferred vote and swing vote from the 2016 election, as well a collection of socio-demographic variables from the 2016 census - also obtained from eechidna. The two-party preferred vote is the percent of votes favouring the Liberal party over the Labor party, and the swing vote is the percentage point change in the two-party preferred vote from 2013 to 2016.

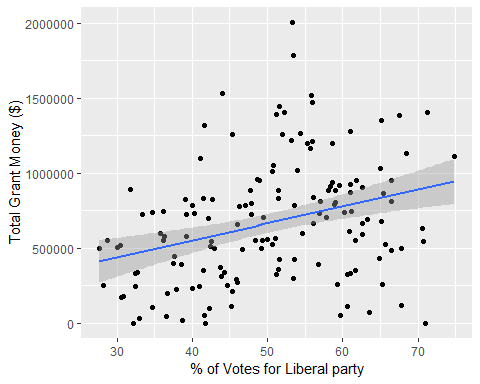
Here’s a snapshot of the resultant dataset.

## # 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.

#### Statistical modelling

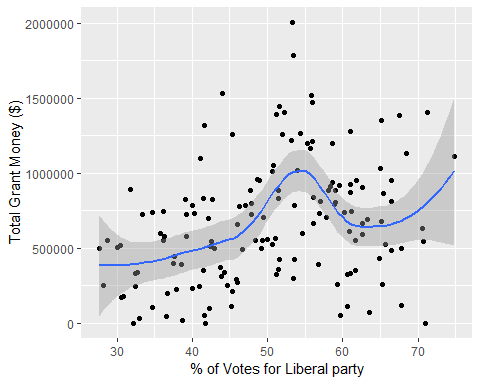
What I’m really looking to uncover is the relationship between electorates’ two-party preferred vote for the Liberal party in the 2016 election and the amount of grant money received. First, let’s see what a scatterplot looks like with a linear regression line fitted.



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

So instead of fitting a linear regression line, I plot a loess curve - a smooth, non-parametric regression curve that gives higher weighting to local data points. This allows for a better picture of the relationship between the 2016 vote and grant amount.

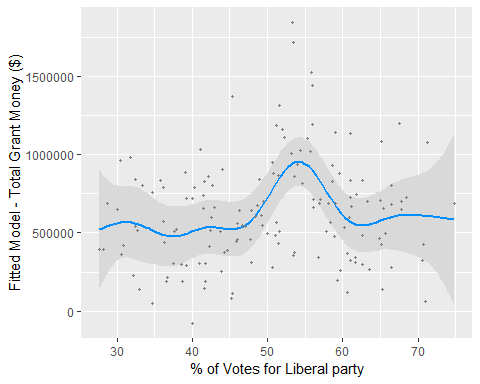
## `geom\_smooth()` using method = 'loess' and formula 'y ~ x'



Wow! The loess curve reveals a pretty significant relationship - one that was not obvious from inspecting the initial scatterplot.

There is a clear spike in the amount of grant money given to electorates that were won by the Liberal party in 2016, but not won by a huge margin. These are called “marginal” Liberal party electorates. This plot in itself is pretty convincing evidence of Liberal bias in the funding scheme.

But to take this another step further, I decide to fit a Generalized Additive Model (GAM) with the grant amount as the response, and the two-party preferred vote (from 2016), swing vote (percentage point change in two-party vote from 2013 to 2016), and census characteristics as covariates. Each of covariate is included in the model as a smooth term. In this case I have use penalized regression splines - a type of flexible non-parametric regression similar to loess.



It can be clearly seen that even after accounting for the swing vote and census variables, there is still a huge spike in grant amount for marginal Liberal party electorates! The p-values from the GAM regression output confirm that this is the conclusive evidence I set out to discover. With a p-value < 0.001, the data proves that the two-party preferred vote had a statistically significant effect on total grant money allocated to each electorate.

Finally, to check that my findings are robust to outliers, I re-estimate the model after removing the two electorates with outstandingly large grant amounts ($1.8M and $2.0M - given to the electorates of Boothby and Dawson), and find that the same Liberal bias is observed. Additionally, the model assumptions seem to be valid, as the residuals are approximately normal (observed using a quantile-quantile plot). Also for those wondering, the fully specified GAM explains 46.3% of the deviance - which I consider to be a pretty good fit in this case.

#### In conclusion

My statistical model provides empirical evidence that the Liberal party did allocate grants with a biased strategy in order to consolidate support for the 2019 election. The data shows that marginal Liberal electorates were targeted by the government and received more grant funds as a result.

While this analysis provides illuminating insight into Australian politics, it also demonstrates how publicly available data can be used to uncover powerful insights.

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) and is available on CRAN.