Who we are affects how we vote

Rob J Hyndman and Di Cook

1 Introduction

We often hear stereotypes about voting patterns — people in their 20s without kids are more likely to be left wing, migrants are more conservative, wealthier people tend to favour the conservative parties, and so on. So we thought we would test these ideas by matching census data to election data, to see if we can identify what are the socio-demographic characteristics of an electorate that are most closely related to how they vote.

It is also interesting to see how this might have changed over time. For example, if wealth was a good predictor of voting patterns in the 2001 election, is it still a good predictor in 2019? Australia has changed in many ways over the last two decades. Rising house prices, country-wide improvements in education, an ageing population, and a decline in religious affiliation, are just some of the ways we have changed. At the same time, political power has moved back and forth between the two major parties. How much can we attribute changes in political power to changes in who we are?

2 Census and electoral data

The Census provides data on electoral socio-demographics, and vote counts in each electorate can be obtained from Australian federal elections. However, joining these two data sources is difficult because the Censuses are not held at the same time as the elections. Between 2001 and 2016 there were six elections and four Censuses, as shown in the timeline below.

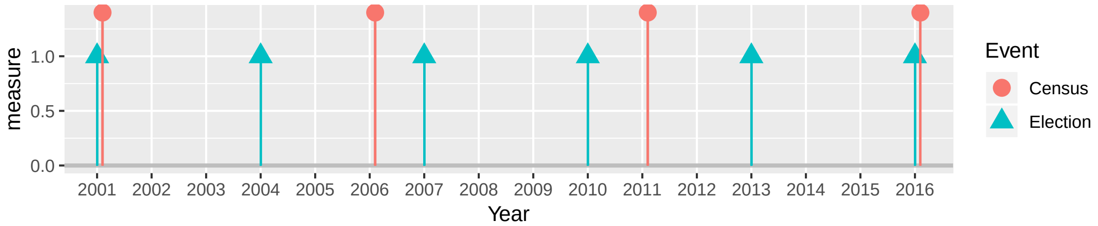


Figure 1: *Timeline of Australian elections and Censuses. They do not always occur in the same year.*

Not only can an electorate change between the last Census and an election, but even the electorate boundaries can change. Some electorates can disappear altogether and new electorates can arise. Electoral boundaries are redistributed regularly by the AEC, meaning that only in the years where both a Census and election occur are all boundaries likely to match — the case for the 2001 and 2016 elections. So we first had to estimate what the socio-demographic characteristics of an electorate would have been at the time of each election using a complicated method of interpolation over time and geography. This method uses Census information from both before and after the election of interest, and information from neighbouring electorates when boundaries have changed.

3 2PP Modelling

A simple way to measure voting patterns is to consider the two-party preferred (2PP) vote, which is based on the tally of preferences for the Labor and Liberal candidates, ignoring all other candidates. By convention, this is recorded as a percentage preference in favour of the Liberal party — for example, a 2PP value of 45% indicates that 45% of voters ranked the Liberal candidate higher than the Labor candidate, while the remaining 55% ranked them in the reverse order.

We consider how various socio-demographic variables obtained from Census data can be used to explain the 2PP values for each of the 150 electorates in each of the federal elections between 2001 and 2016.

Many of the socio-demographic variables have changing scales over the years. For example, inflation-adjusted median rental prices increased across almost all electorates, with median rent of 200 dollars per week placing an electorate in the 90th percentile in 2001, but only the 30th percentile in 2016. In order for socio-demographic effects to be comparable across years, all socio-demographic variables were standardized.

There are dozens of socio-demographic variables available in the Censuses, with many variables representing similar information about an electorate. So we combined some variables to avoid redundant information. For example, our "Incomes" variable is a combination of median personal income, household income and family income.

Each election was modelled separately, to allow us to see any changes over time, and to account for changing electorate boundaries. In this article, we highlight the variables with the strongest relationship to the two-party preferred vote, or which have had substantial changes over time. The full analysis is available at https://robjhyndman.com/publications/elections/.

3.1 Country-wide trend

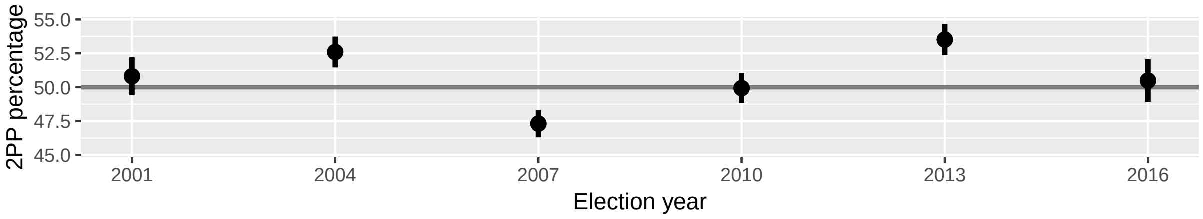


Figure 2: *Estimated intercept for each election, which represents the two-party preferred vote for an electorate with average characteristics.*

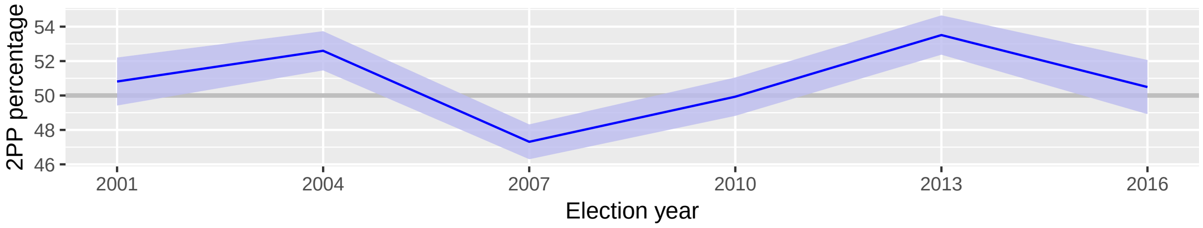
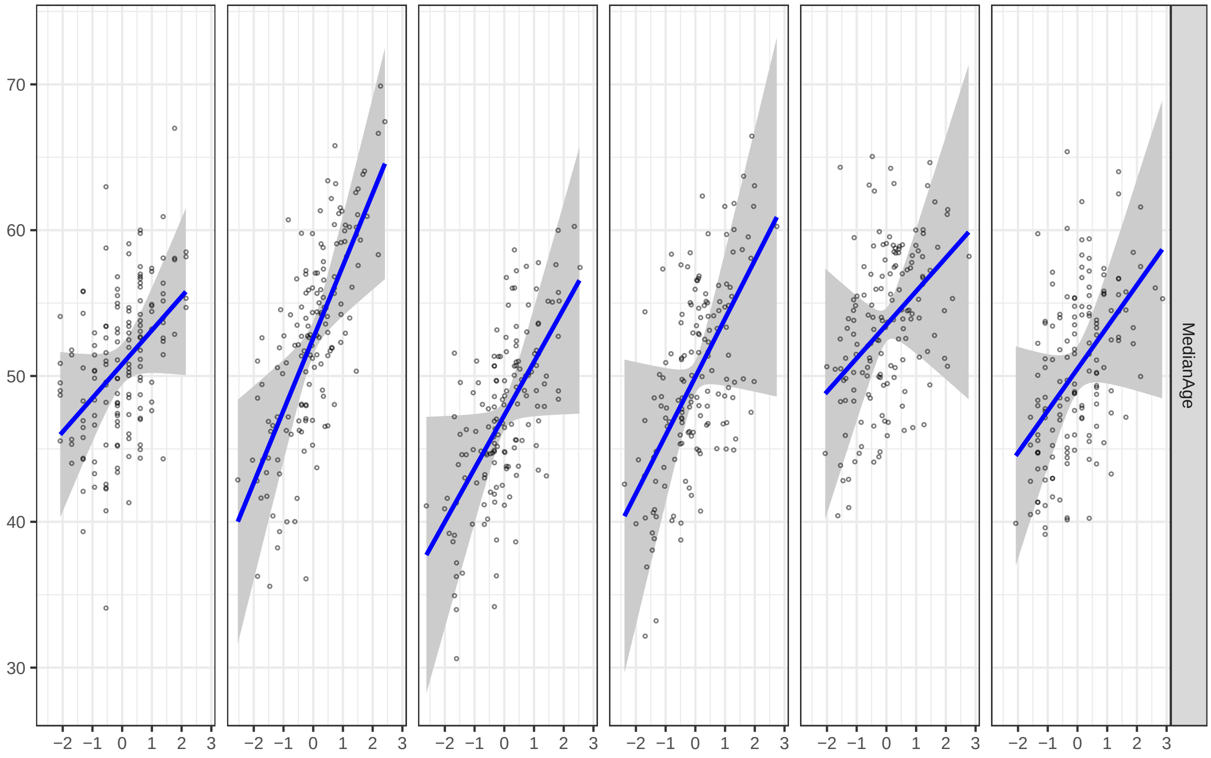


Figure 3: *Estimated intercept for each election, which represents the two-party preferred vote for an electorate with average characteristics.*

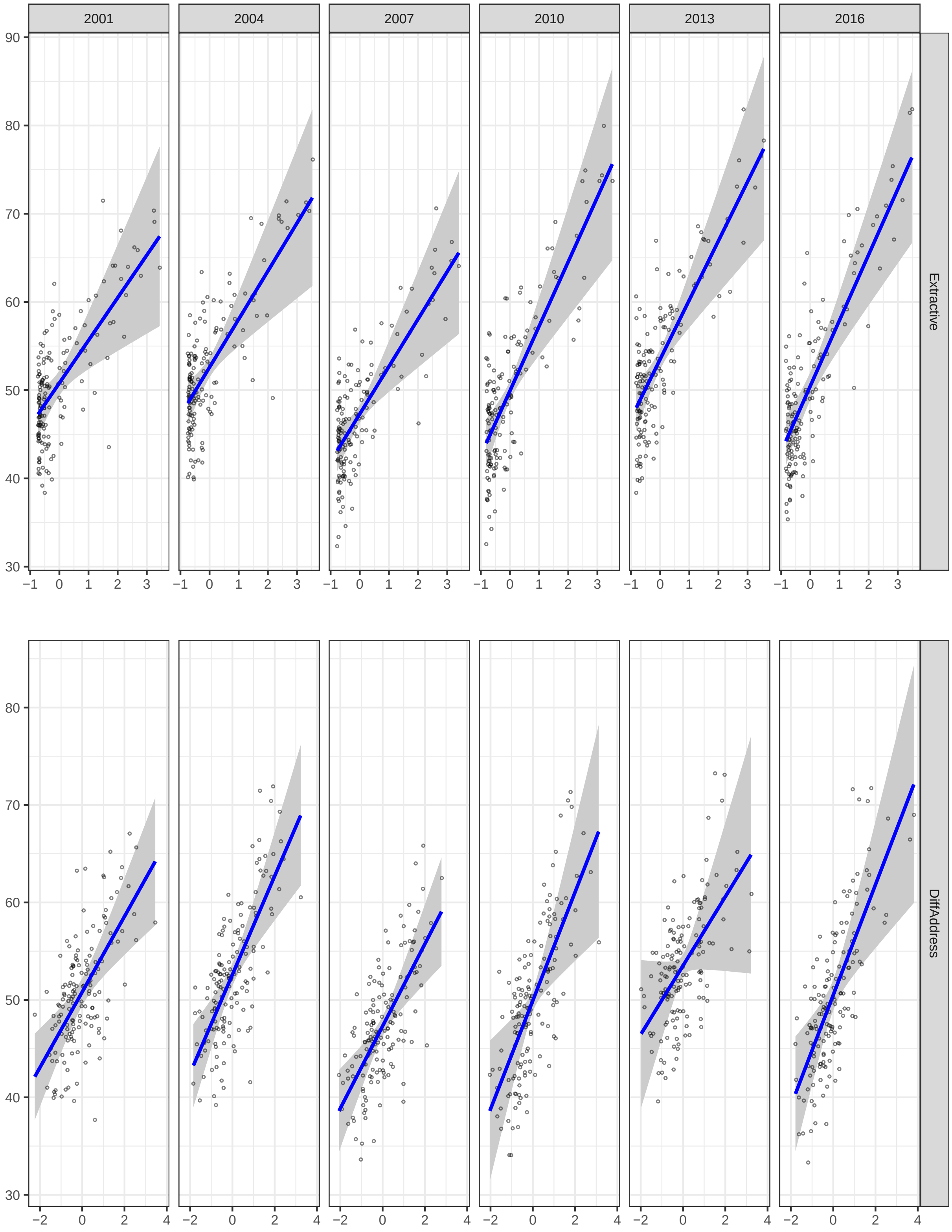
Since all socio-demographics have been standardized, the intercept in each model can be interpreted as the estimated two-party preferred vote for an "average" electorate. Figure 2 shows that the baseline of party preference has varied over the elections, with the biggest swing occurring in the 2007 election where the mean electorate shifted more than five percentage points in favour of the Labor party. The dots represent the estimated 2PP value, and the lines indicate a 95% confidence interval providing a guide to the uncertainty in the estimate.

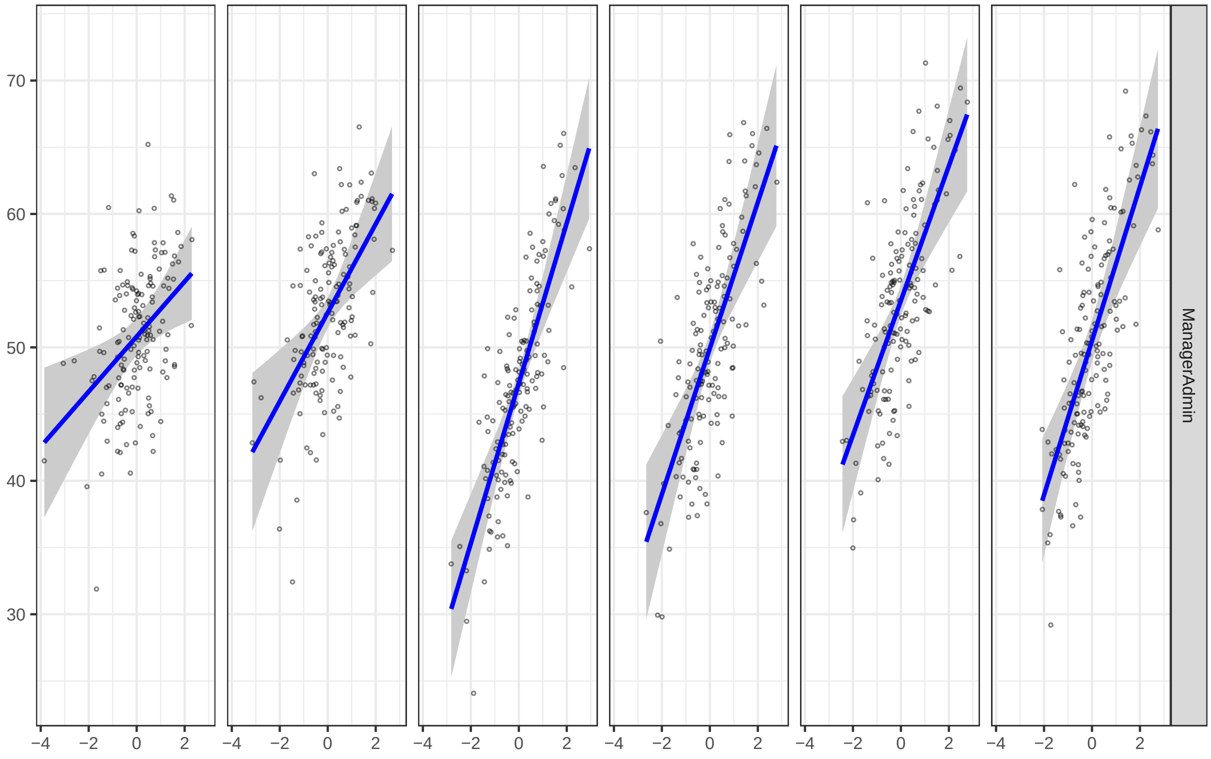
3.2 Age

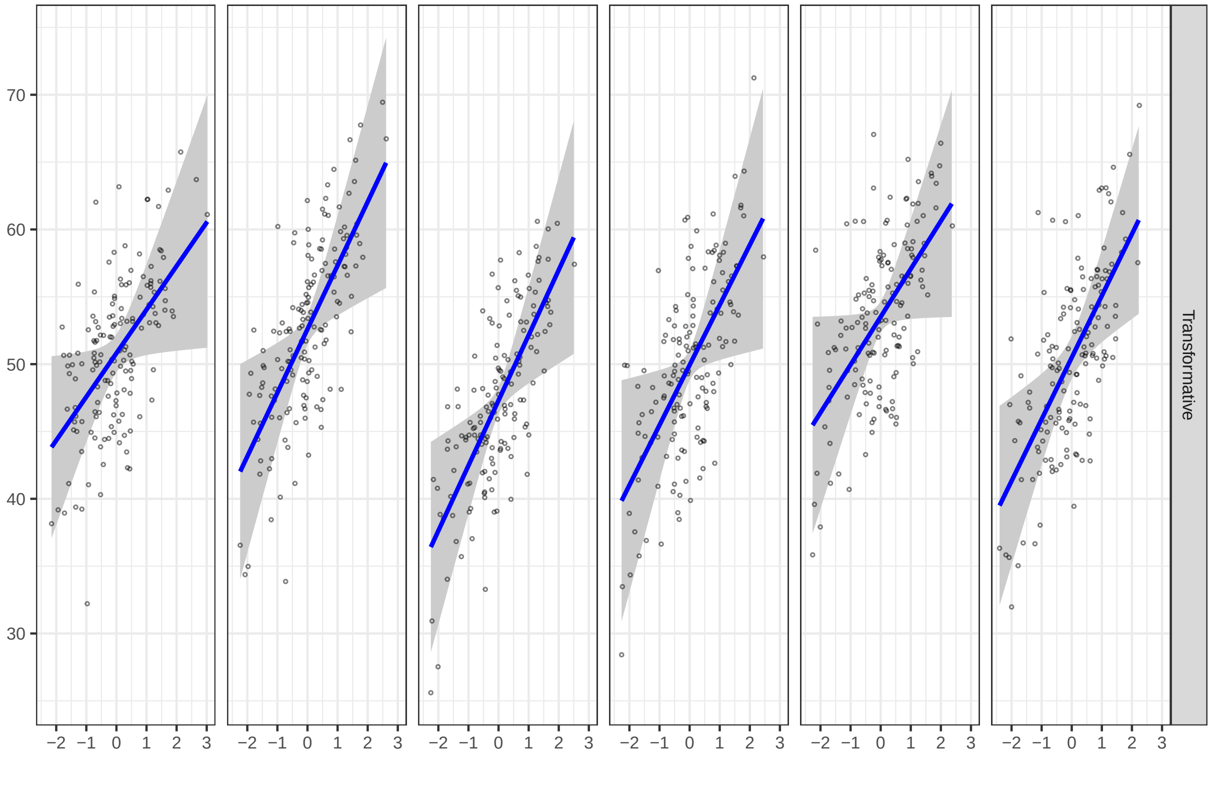
- Would it be possible to show how the correlation changes for different age brackets (ie. younger voters = more Labor, and this swings more to Liberal as the age brackets gets older)?
- If not, can you go into more detail specific age brackets? Eg. Is it the over-65s that skew Liberal?

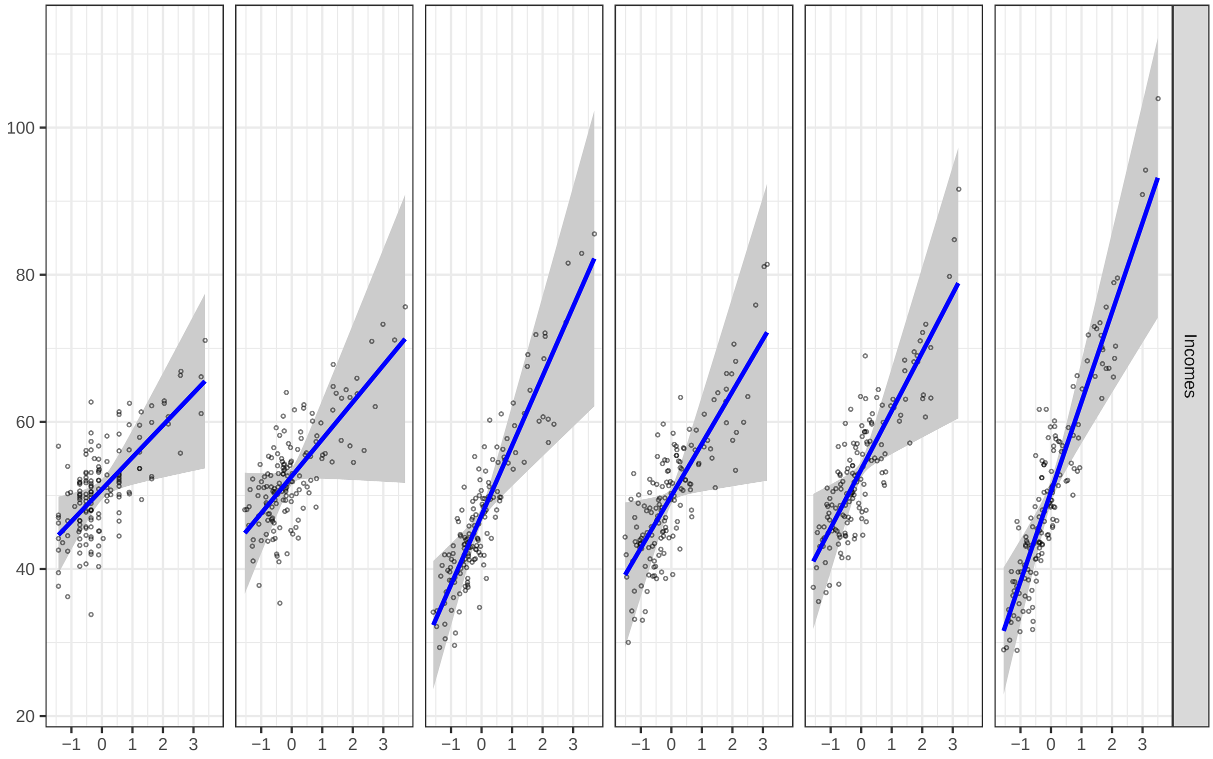


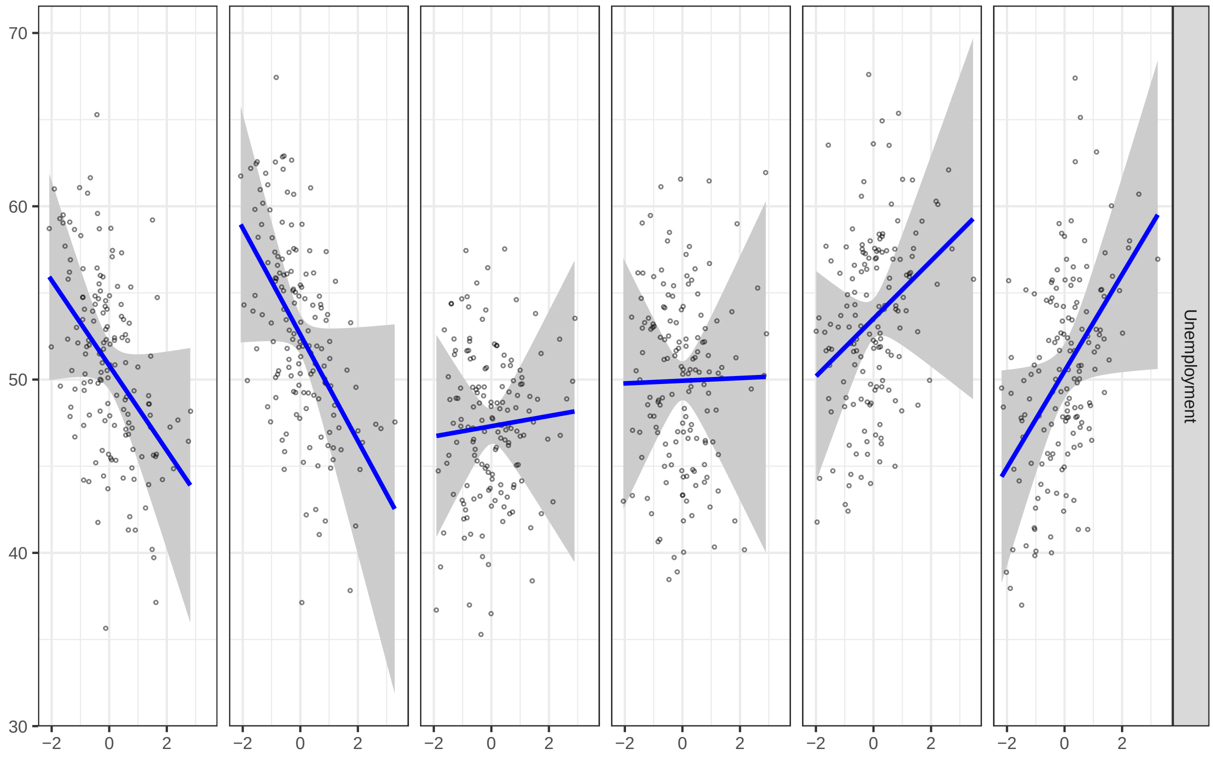
3.3 Influential socio-demographics

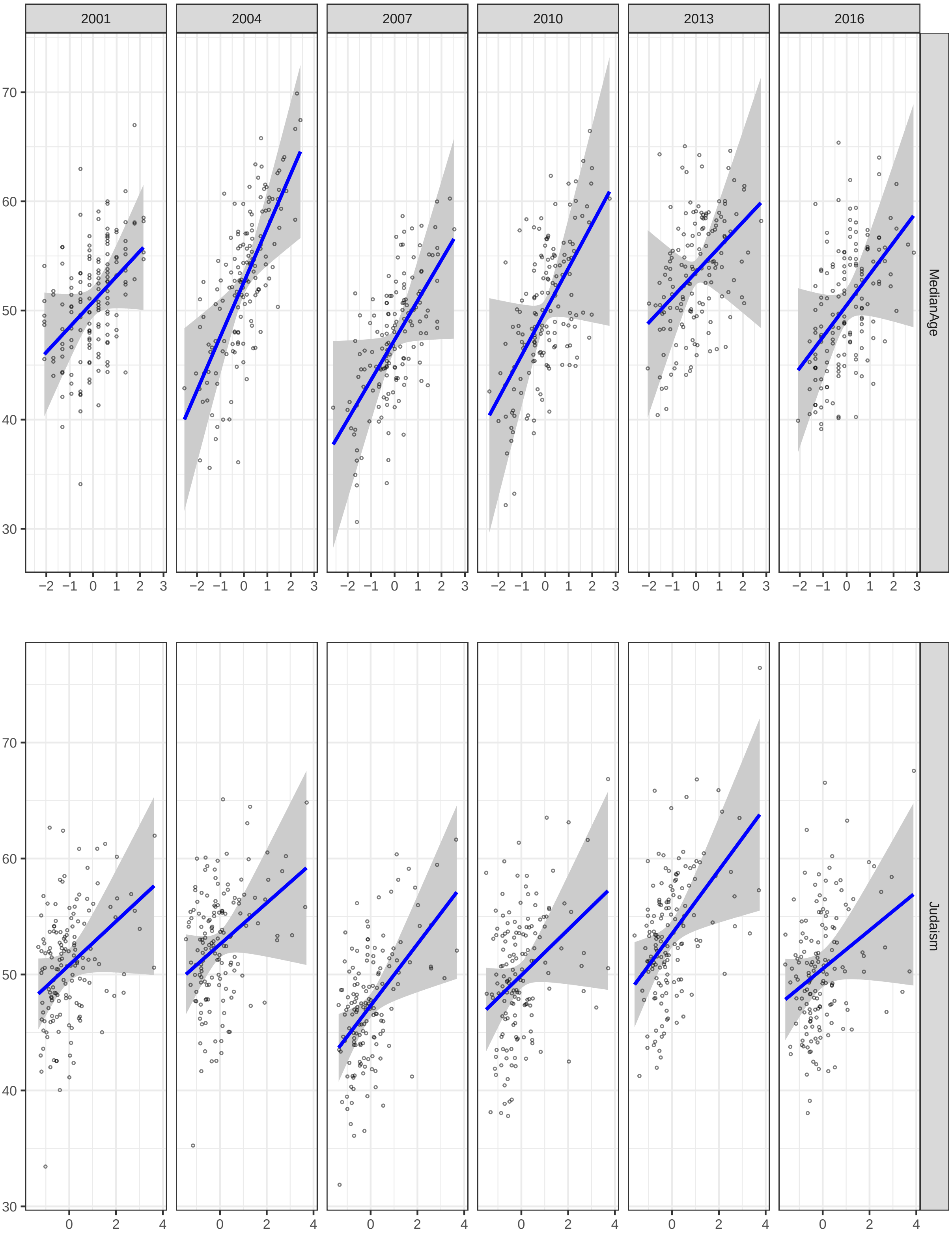


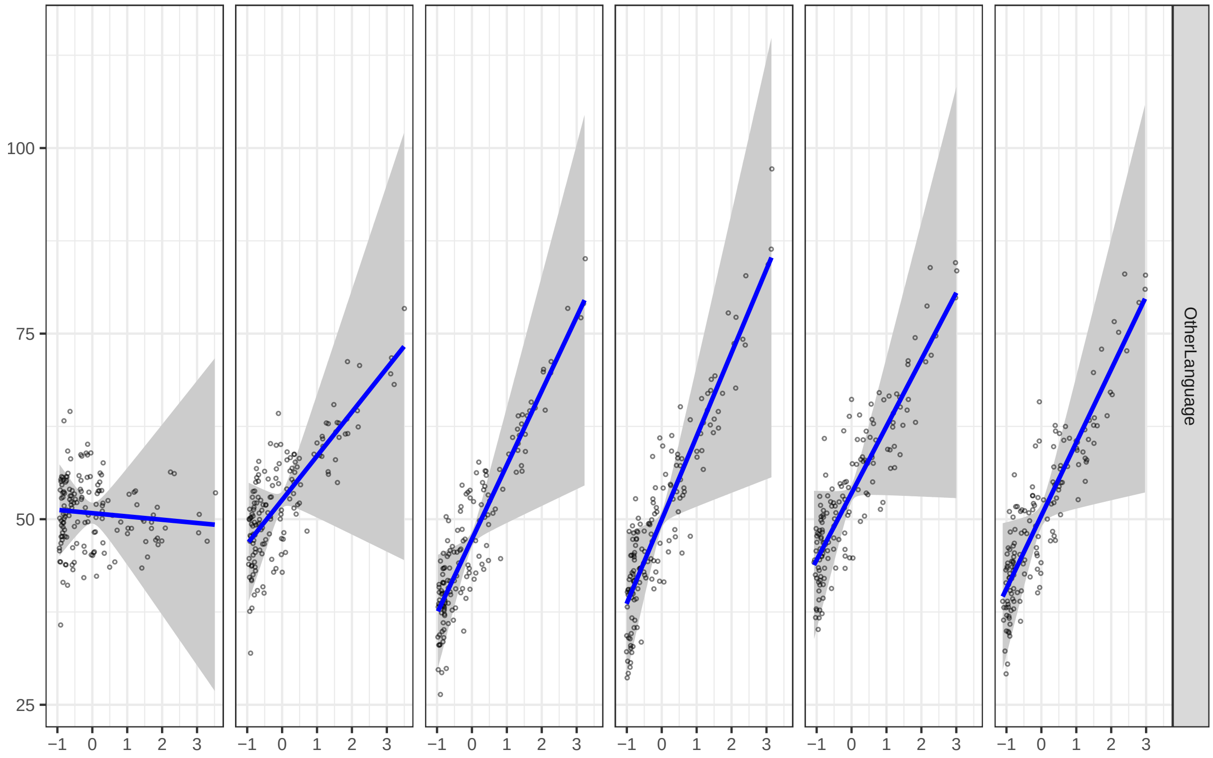

70
60
50
40
30
−4 −2 0 2
−4 −2 0 2
−4 −2 0 2
−4 −2 0 2
−4 −2 0 2
−4 −2 0 2
ManagerAdmin

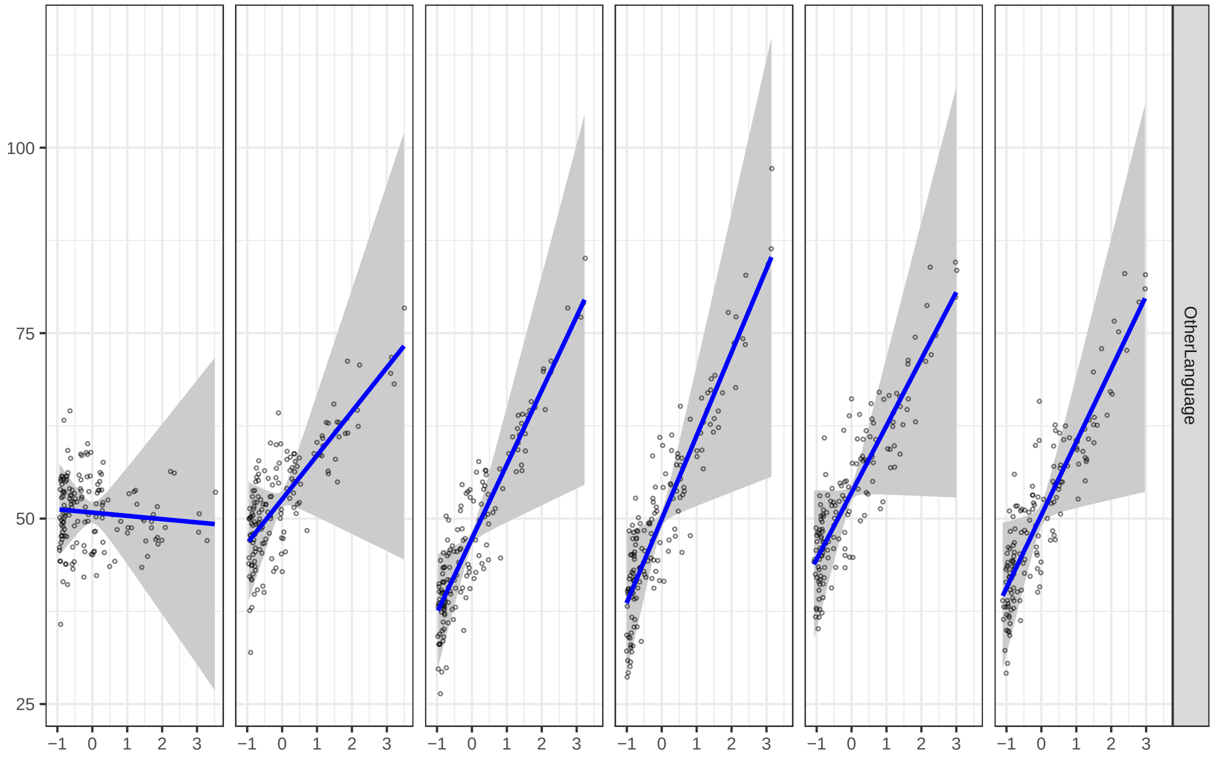

70
60
50
40
30
−2 −1 0 1 2 3
−2 −1 0 1 2 3
−2 −1 0 1 2 3
−2 −1 0 1 2 3
−2 −1 0 1 2 3
−2 −1 0 1 2 3
Transformative

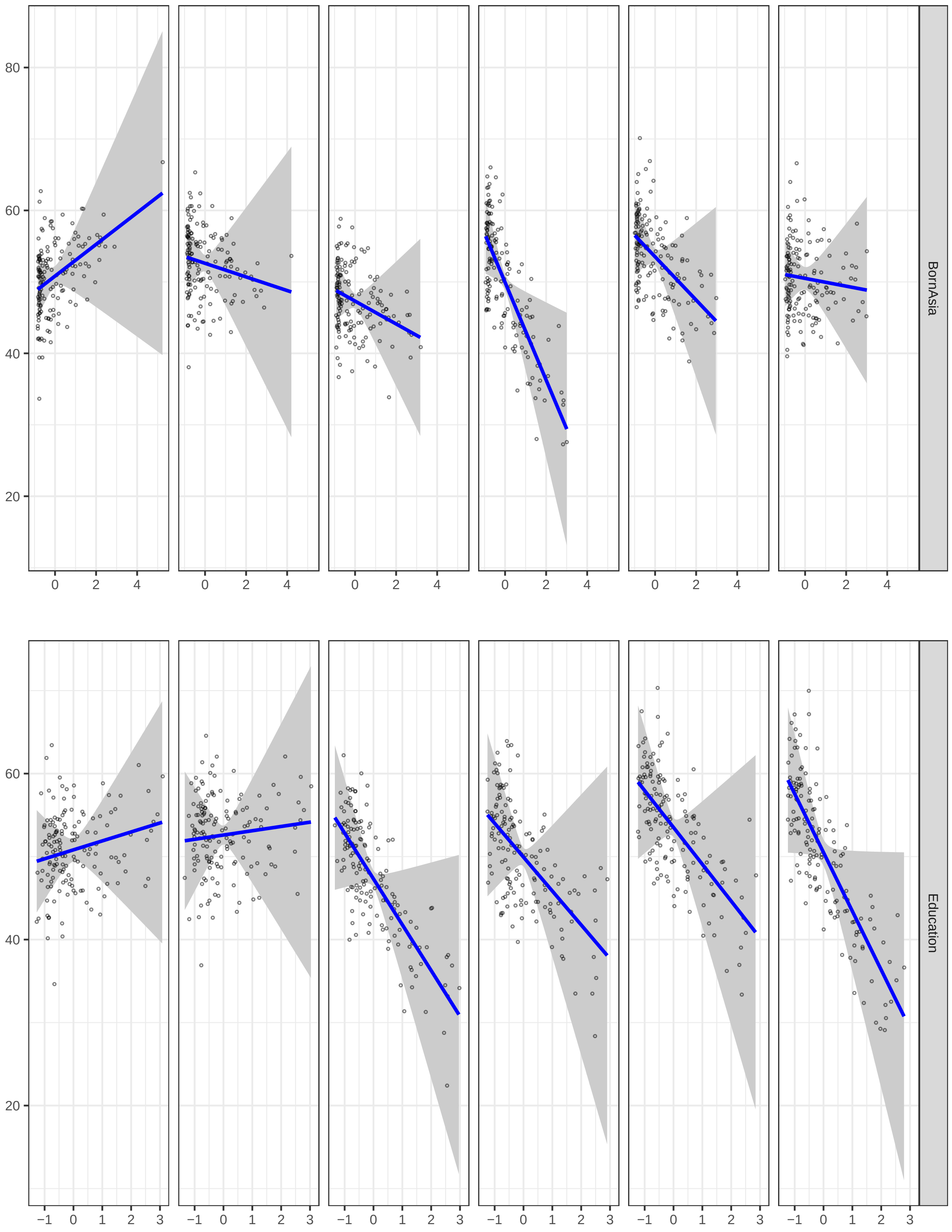

100
80
60
40
20
−1
0
1
2
3
Incomes

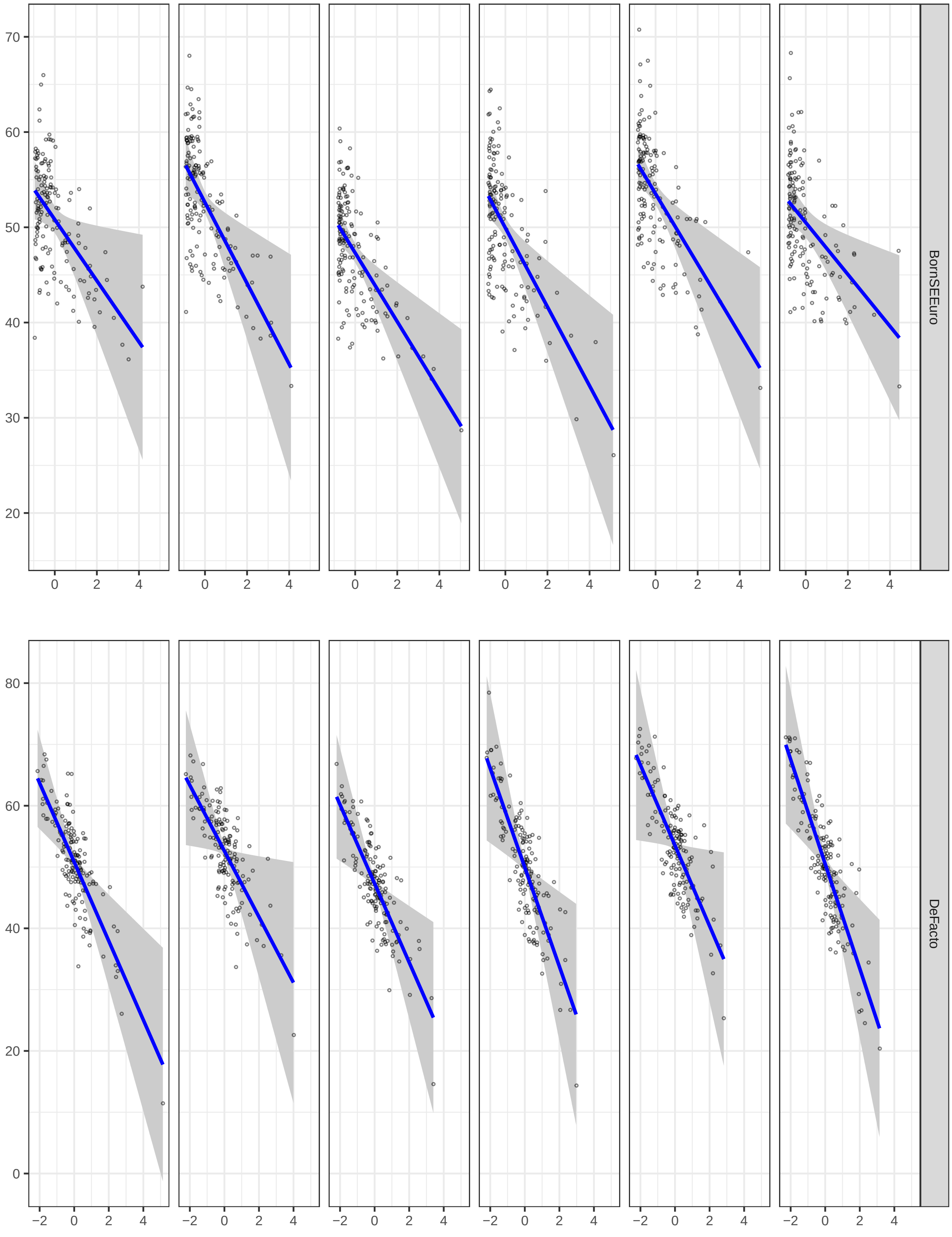

70
60
50
40
30
−2
0
2
Unemployment


2001
2004
2007
2010
2013
2016
MedianAge
Judaism
70
60
50
40
30
−2
−1
0
1
2
3
0
2
4


100
75
50
25
−1
0
1
2
3
OtherLanguage


100
75
50
25
−1
0
1
2
3
OtherLanguage


80
60
40
20
0
2
4
BornAsia
60
40
20
−1
0
1
2
3
Education



Partial residual plots by election year for a selection of predictors. Linear model with 95\% confidence bands overlaid. Most predictors have a positive relationship: the larger the value the more likely the electorate preferences the Coalition. The relationship is relatively robust over time, with the exception of Unemployment.

Several predictors have a negative relationship: with larger values indicating the electorate more likely preferences Labor. Most relationships are relatively stable over elections, except

OtherLanguage and Education.

To investigate the socio-demographics that have a strong effect on the two-party preferred vote, partial residual plots are used and shown in Figures **??** and **??**. The partial residuals are the residuals from the fitted model with the estimated effect an individual variable added. These show the direction, size and significance of an estimated effect — the slope of the prediction line matches the estimated coefficient, and the shaded region represents a 95% confidence band, computed using the method in Breheny & Burchett (2017). If a horizontal line can be drawn through the confidence band, then the effect is insignificant. The estimated intercept is also added to the partial residuals for interpretability. Plots for each election are faceted to compare the effects over time in Figures **??** and **??**. Only socio-demographics that have a significant effect in at least one election are displayed in Figures **??** and **??**.

It is important here to note the ecological fallacy: insights are being drawn at the electorate level, and cannot be inferred for another disaggregate level (e.g. individual voters).

3.3.1 Income and unemployment

Typically the Labor party campaigns on more progressive policies, which often include tax reform that adversely affects higher income earners, and more generous social assistance programs. Perhaps it is due to these policies that higher income electorates appear more likely to support the Liberal party, as the `Incomes` factor has a positive effect on Liberal preference (see row 1 in Figure **??**). This effect is significant in every election aside from 2004, where it is only marginally insignificant ($p = 0.0613$). Unemployment however, is not as influential. In 2001 and 2004, electorates with higher unemployment align with Labor, but over time this shifts towards support for the Liberal party, culminating in a significantly positive effect in 2016.

3.3.2 Industry and type of work

Electorates with higher proportions of workers in mining, gas, water, agriculture, waste and electricity (grouped as `Extractive` industries) are consistently linked with higher support for the Liberal party, with the magnitude of this effect slightly increasing over the years (see row 3 in Figure **??**). This is unsurprising, as the Liberal party has close ties with these traditional energy industries, and typically present policies to reduce taxation on energy production. Furthermore, electorates with more workers in construction or manufacturing industries (`Transformative`) are also more likely to support the Liberal party (see row 4 in Figure **??**).

Similarly, the proportion of workers in managerial, administrative, clerical and sales roles (`ManagerAdmin`) is also a significant predictor of two-party preference vote across all six elections, with a higher proportion of people working these jobs increasing Liberal support. The magnitude of this effect also seems to increase over the years.

3.3.3 Household mobility

In each of the six elections, electorates with a higher proportion of people that have recently (in the past five years) moved house (`DiffAddress`) are more likely to support the Liberal party, although this effect was marginally insignificant in 2013 (see row 6 in Figure **??**. Having controlled for characteristics of house ownership and rental prices (via the factors `PropertyOwned` and `RentLoan` respectively), this effect is somewhat surprising.

3.3.4 Relationships

De facto relationships, but not marriages, are found to be an important (and significant) predictor of the two-party preferred vote in all six elections, with more de facto relationships associated with higher support for the Labor party. The proportion of individuals who are married however, is insignificant (not shown).

3.3.5 Age

Regions comprising more older people are often believed to be more conservative, and indeed it found that electorates with a higher median age are more likely to support the Liberal party — although this effect is significant only in 2007 and 2010 (see row 2 in Figure **??**).

3.3.6 Education

Since 2007, electorates with higher education levels are associated with supporting the Labor party, although this effect is significant only in 2016. Before 2007, education has an almost zero effect (see row 3 in Figure **??**).

3.3.7 Diversity

Larger migrant populations from Asia, the Middle East, South-Eastern Europe, the United Kingdom and elsewhere, are either associated with Labor support, or have no effect. Of these areas, only South-Eastern European populations appear significant in each election, with the proportion of Asian migrants also being significant in 2010. Speaking other languages (aside from English) however, appears to have a far stronger effect, as observed through the `OtherLanguage` variable. Electorates with more diverse speech are associated with higher support for the Liberal party from 2004 onwards, with this effect being significant in 2007, 2010 and 2016. Furthermore, of the variables relating to religion, only Judaism shows a consistent effect, with electorates with relatively large Jewish populations more likely to vote Liberal.

3.4 A closer look at the residuals

3.4.1 Residuals by state

It is often hypothesized that states have systematic differences that cause their electorates to vote differently. Boxplots of residuals grouped by state (Figure 4) reveal that the data reflects this – there appears to be a state-specific effect not captured by the models. Tasmania and the Australian Capital Territory appear to have a bias towards Labor, whereas the Northern Territory tends towards voting Liberal. However, there are relatively few electorates in each of these states (five, two and two respectively), so this apparent result may be due to incumbent effects rather than an actual state-specific bias.

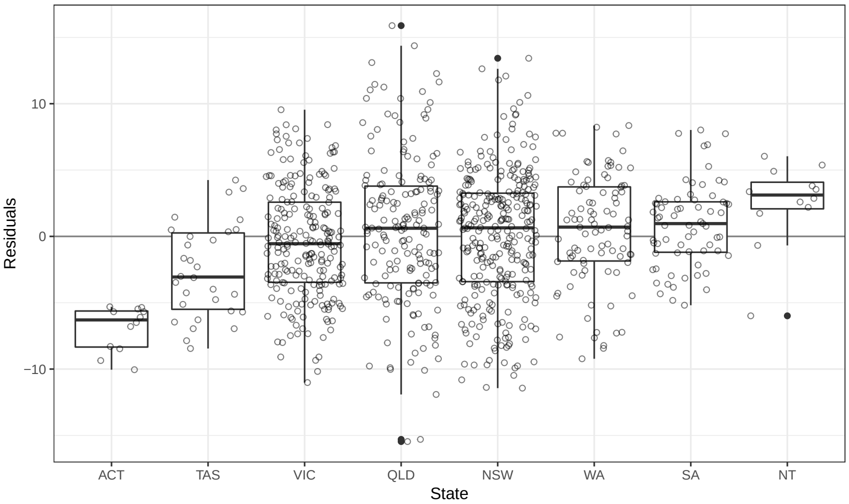


Figure 4: *Boxplot of residuals by state with jittered points. States ordered by median residual. A state-specific bias not captured by the model is evident.*

3.4.2 Outlier electorates

Based on the distribution of the Cook's distance values, a Cook's distance greater than 0.1 is considered to be influential and a potential outlier. The electorate of Sydney (NSW) has a large Cook's distance from 2001 to 2013, due to its diverse population (language, birthplace and religion), high number of defacto relationships, high income, high household mobility and small amount of workers in extractive and transformative jobs. It has remained a strong supporter of the Labor party and the Liberal vote is severely overpredicted by the model, making it an outlier. Nearby in metropolitan NSW, the electorate of Wentworth is found to be an outlier in all but the 2007 election. Although historically Liberal, its two-party vote jumped by over 10 percentage points in 2010 without experiencing any notable changes in its socio-demographic makeup — implying that this may be the direct effect of its Liberal member, Malcolm Turnbull, becoming the leader of the Liberal party. Liberal support in Wentworth is underpredicted by the model in each year, and more so with Turnbull as Liberal leader.

Lingiari, an electorate taking up almost all of the Northern Territory, is an outlier in the 2001–2007 elections due to its large Indigenous population, young age profile and low rates of property ownership. Fowler (NSW) has a diverse population with a high proportion of migrants, many Buddhists and Muslims, and has strong Labor support, making it influential in 2001, 2004 and 2010. Other electorates with large Cook's distance are Barton (NSW) and Leichhardt (QLD) in 2016, and Canberra (ACT) in 2007.

4 Conclusion

This paper explores the effects of electoral socio-demographic characteristics on the two-party preferred vote in the 2001–2016 elections, using information from the corresponding Australian federal elections and Censuses. As a Census does not always occur in the same year as an election, Census data for the 2004–2013 elections are generated by employing a method of spatio-temporal imputation. This imputes electoral socio-demographics for the electoral boundaries in place at the time of the election — an approach that is distinctly different from previous work on modelling election outcomes, where Census and election data are typically joined without addressing their temporal differences. Before estimating a model, these socio-demographic variables are standardized (to adjust for changing variable scales) and many variables (representing similar information) are combined into factors, resulting in a reduced predictor set. A spatial error model is then estimated for each election, accounting for the inherent spatial structure of the data.

Across the past six elections, most of the socio-demographics that drive the electoral two-party preferred vote are found to remain steady, whilst a few (typically weaker) effects vary over time. Industry and type of work are particularly influential, with energy-related and manufacturing/construction jobs, as well as administrative roles being strongly linked with the Liberal party in all elections. Incomes have a similarly consistent effect, with higher income areas supporting Liberal. Higher levels of unemployment shift from weak association with Labor to a significant Liberal effect over the years, and higher education levels are associated with Labor from 2007 (although significant only in 2016). It is also found that electorates with higher household mobility support Liberal, birthplace diversity favours Labor and more de facto relationships align with Labor preference — although marriages, family and household sizes have no material influence. Furthermore, the neighbourhood (spatial) effects are found to be positive in all elections, although significant only in 2001 and 2016, meaning that in the 2004–2013 elections, electorates effectively voted independently.

The findings in this paper complement the existing literature by modelling temporal trends, which as far as the authors are aware, has not been done previously for Australian elections using a regression framework. It is also the first study to model any Australian election since 2010 using Census information.

Additionally, a key contribution of this research is the wrangling of the raw data and imputed data sets for the 2004, 2007, 2010 and 2013 elections, which have been contributed to the `eechidna` `R` package — providing a rich, accessible data resource for future Australian electoral analysis.

5 Acknowledgements

This paper was produced using `RMarkdown` (Allaire et al. 2019) and `knitr` (Xie 2015). All corresponding code for this paper can be found in the github repository github.com/jforbes14/eechidna-paper, and the data used is available in the `eechidna` package (Forbes et al. 2019). All raw data was obtained from the Australian Electoral Commission, the Australian Bureau of Statistics and the Australian Government.

6 Software

All election and Census datasets, along with electoral maps and more, are available in the `eechidna` (Exploring Election and Census Highly Informative Data Nationally for Australia) `R` package, which can be downloaded from CRAN. The `eechidna` package makes it easy to look at the data from the Australian Federal elections and Censuses that occurred between 2001 and 2016. This study contributed a large revision to the eechidna package, which included the addition of election and Census data for 2001–2010, voting outcomes for polling booths and imputed Census data for election years. For more details on using eechidna, please see the articles (vignettes) on the github page ropenscilabs.github.io/eechidna/.

The authors would like to sincerely thank Anthony Ebert, Heike Hofmann, Thomas Lumley, Ben Marwick, Carson Sievert, Mingzhu Sun, Dilini Talagala, Nicholas Tierney, Nathaniel Tomasetti, Earo Wang and Fang Zhou, all of whom have contributed to the eechidna package.

7 From Emil

In terms of structure, start with a brief paragraph or two on what the research is, and what the reader can expect to find below.

Then, break the article into small sections for each of the factors that have an effect (a short paragraph or two for each factor). And it would be great to include examples of specific seats where the relationship was strong, or interesting for other reasons.

So the sections will be:

- Age
- Income
- Unemployment
- Industry and type of work
- Education
- Diversity
- Household mobility
- Relationships

An example of this structure can be seen here: https://theconversation.com/confused-about-aged-care-in-the-ho

Questions on particular factors:

Incomes: - Was this correlation stronger in 2016? (and increasingly influential?)

Household mobility - In the report, you say: "“Having controlled for characteristics of house ownership and rental prices”, whis effect is somewhat surprising”. Why is this surprising to you? (I would have expected this to support Labor more)

Age - Would it be possible to show how the correlation changes for different age brackets (ie. younger voters = more Labor, and this swings more to Liberal as the age brackets gets older)? - If not, can you go into more detail specific age brackets? Eg. Is it the over-65s that skew Liberal?

Diversity - Migrant population: Would it possible to break this out into particular groups? eg. Seats with higher Chinese population correlates with Liberal, seats with higher SEAsia pop correlates with Labor, etc? - The same for languages?

References

Allaire, J, Y Xie, J McPherson, J Luraschi, K Ushey, A Atkins, H Wickham, J Cheng, W Chang & R Iannone (2019). *rmarkdown: Dynamic Documents for R*. R package version 1.12. `https://rmarkdown.rstudio.com`.

Breheny, P & W Burchett (2017). Visualization of regression models using visreg. *The R Journal* **9**(2), 56–71. `https://journal.r-project.org/archive/2017/RJ-2017-046/index.html`.

Forbes, J, D Cook, A Ebert, H Hofmann, RJ Hyndman, T Lumley, B Marwick, C Sievert, M Sun, D Talagala, N Tierney, N Tomasetti, E Wang & F Zhou (2019). *eechidna: Exploring Election and Census Highly Informative Data Nationally for Australia*. R package version 1.3.0. `https://CRAN.R-project.org/package=eechidna`.

Xie, Y (2015). *Dynamic Documents with R and knitr*. 2nd ed. Boca Raton, Florida: Chapman and Hall/CRC. `http://yihui.name/knitr/`.