Model-Based Pre-Election Polling for National and Sub-National Outcomes in the US and UK

Benjamin Lauderdale

31 October 2017

London School of Economics and Political Science

Intro

OUTLINE OF TALK

1. Introduction
2. Weight-Based Polling Methods
3. Model-Based Polling Methods
    - Decomposition of the problem
    - Data sources
4. Results
    - UK Referendum on EU Membership, June 2016
    - US Presidential Election, November 2016
    - UK General Election, June 2017
5. Discussion

This talk is based on a collaboration with Doug Rivers (YouGov/Stanford), Delia Bailey (YouGov), and Jack Blumenau (UCL).

1. UK referendum on EU membership (23 June 2016)
    - We did a limited public release of the findings before the election. (https://yougov.co.uk/news/2016/06/21/yougov-referendum-model/)
    - Final estimate was Leave 50.6 (ahead by 1.2%) versus actual result of 51.9% (ahead by 3.8%). Error of 2.6% on margin.
2. US presidential election (8 November 2016)
    - We posted daily updates of our estimates for the month preceding the election (https://today.yougov.com/us-election/)
    - Final estimate for popular vote was Clinton +3.8% versus actual +2.1%. Error of 1.7% on margin.
3. UK general election (8 June 2017)
    - Daily updates of our estimates for the eight days preceding the election. (https://yougov.co.uk/uk-general-election-2017/)
    - Final estimate for popular vote was Conservatives +3.4% versus actual +2.5%. Error of 0.9% on margin.

Poll firm predicts shock losses for Theresa May's Tories at general election

Controversial YouGov estimate points to hung parliament with 20 fewer seats for May

YouGov's poll predicting a hung parliament is certainly brave

The polling firm is employing a new 'controversial' methodology only 10 days before the general election

POLLS APART Shock new poll shows Theresa May LOSING her majority – but experts brand it 'utter tripe'

> The survey, which contradicts every other poll, suggests that the Tories will win 310 seats

Figure 1: Impact Case Study?

Weight-Based Polling Methods

HOW ARE MOST POLLS CONDUCTED?

- In theory:
    1. Collect a sample by telephone or online.
    2. Calculate population weights $\hat{W}_i(X_i)$ to make sample match known distribution of eligible voters on measured covariates
    3. Calculate turnout probability weights $\hat{T}_i$ from stated intention to turnout and/or other variables, for each sampled respondent
    4. National vote share for party k is

$$\frac{\sum_i V_{ik} \hat{T}_i \hat{W}_i(X_i)}{\sum_i \hat{T}_i \hat{W}_i(X_i)}$$

- In practice...

PROBLEMS WITH SAMPLES

- Good samples are difficult to come by
    - Almost none of the pre-election polls even pretend to be probability samples.
- Non-probability samples are full of weird people
    - People who answer their phones are unusual.
    - People who fill out online surveys are unusual.
- Common biases:
    - Too few young people
    - Too few people with lower educational attainment
    - Too few people who are not interested in politics.

PROBLEMS WITH POPULATION WEIGHTING

- Generally, the variables for which you have sound targets (census, election results) do not include all known sample selection biases (esp political interest / attention).
    - Some pollsters use dubious targets
- You have to weight to the voting *eligible* population, which means if you weight on lagged vote choice, you need to include non-voters.
    - Some pollsters do this incorrectly

PROBLEMS WITH TURNOUT MODELLING

- Generally the implied turnout rate from self-reported turnout/vote intention, after the above weighting, is still way too high.
    - Could be because of turnout overstatement
    - Could be because the samples are unrepresentative **even after** weighting
    - Pollsters engage in a range of atheoretical tweaks at this stage, or just ignore the problem
- Applying a demographic turnout filter at this point is not theoretically sound
    - Some pollsters do this incorrectly

LIMITATIONS

- Fundamental problems
  - The quality of samples is low, in ways that are difficult to fix.
  - Not much good for sub-national estimates, and thus for seat projections in non-PR systems.
  - It is difficult to evaluate and refine your methods on the basis of a single national vote share estimate once every few years.
- Obviously the solution is to make the analysis much more ambitious!

Model-Based Polling Methods

WHAT IS MODEL-BASED POLLING?

- Weight-based methods estimate weights $\hat{W}_i(x_i)$ as a function of demographics that are applied and summed across the sampled observations i to form vote share estimates:

$$\frac{\sum_i V_{ik}\hat{T}_i\hat{W}_i(X_i)}{\sum_i \hat{T}_i\hat{W}_i(X_i)}$$

- Model-based methods use the sample to estimate a vote choice & turnout model as a function of demographics, which is then fitted and summed across voter types X_i in the target population to form vote share estimates.

$$\frac{\sum_{X_i} \hat{p}(V_i = k|X_i)\hat{p}(T_i = 1|X_i)f(X_i)}{\sum_{X_i} \hat{p}(T_i = 1|X_i)f(X_i)}$$

- Model-based analyses of polling data generally employ some variant on multilevel regression and post-stratification (MRP)
    - The "multilevel" is not required, but you do need the regression to estimate $\hat{p}(V_i|X_i)$ and the post-stratification is the summing over $f(X_i)$.

APPLICATIONS OF MRP TO PRE-ELECTION POLLING

- 2012 US Presidential Election
    - Xbox study (Wang, Rothschild, Goel, and Gelman 2014)
- 2016 US Presidential Election - Florida
    - NYTimes/Upshot Replication Exercise (Corbett-Davies, Gelman and Rothschild)
- 2017 UK General Election
    - Lord Ashcroft (unnamed minions)

- A 'voter type' is a set of measurable characteristics for an eligible voter X_i (age, gender, education, how they voted in the preceding election, where they live, etc). For each voter type, there are three important quantities that we would like to know...
    1. $p(V_i|X_i, T_i = 1)$ What proportion of each type will vote for each electoral alternative among those who do vote?
    2. $p(T_i = 1|X_i)$ What proportion of each voter type will turn out to vote?
    3. $f(X_i)$ or $p(X_i)$ How many individuals in or what proportion of the electorate is of each type?

ESTIMATING $p(v_i|x_i, t_i = 1)$

What proportion of each type will vote for each electoral alternative, among those who do vote?

- Model conditional vote choices of different types as well as possible.
- Multilevel binary or multinomial logistic regression with *many* interactions.
    - Pure prediction problem, use priors/regularization to avoid over-fitting.
- We used YouGov's large, opt-in online panel, but one could use smaller polls.
- Key assumption: you can condition on enough stuff to estimate correct conditional probabilities of supporting each voting alternative, in spite of whatever problems there might be with the representativeness of your sample.

ESTIMATING $p(t_i = 1|x_i)$

What proportion of each voter type will turn out to vote?

- Model conditional probability of turning out for different types as well as possible.
- Multilevel binary logistic regression with many interactions.
    - Pure prediction problem, use regularization to avoid over-fitting.
- We used high quality face to face surveys from the preceding general elections.
- Key assumption: we cannot reliably predict changes in relative turnout of different groups. Given historical stability of turnout, it is better to have a high quality estimate of past turnout patterns than a low quality estimate of future turnout patterns.

ESTIMATING $p(x_i)$

What proportion of the electorate is of each type?

- Construct best feasible joint distribution of types within voting eligible population
    - Must include all the variables in the voting and turnout models
    - May require imputation / reweighting using several data sources
    - Census microdata samples are sensible starting point
- We kept these in the form of large pseudo-samples, about 200k-2m weighted observations
    - Construct fitted values for each pseudo-obs, then aggregate as desired
- Key assumption: you can get enough variables onto this *poststratification frame* accurately enough to vindicate the assumptions on the previous slides with respect to the conditional vote and turnout being close to correct.

DATA FOR ESTIMATING $p(v_i|x_i, t_i = 1)$

What proportion of each type will vote for each voting alternative, among those who do vote?

- EU 2016
  - YouGov UK panel, 2-4k responses, per day, nationally. 14 day data window, with time trends modeled.
  - Key variables: parliamentary constituency, age, qualifications, social grade, gender, and 2015 vote.
- US 2016
  - YouGov US panel, 5-10k responses, per day, nationally. 14 day data window, with time trends modeled.
  - Key variables: congressional district within state, race, age, education, marital status, gender, and 2012 vote.
- UK 2017
  - YouGov UK panel, 7-9k responses, per day, nationally. 7 day data window, with time trends modeled.
  - Key variables: parliamentary constituency, age, qualifications, political attention, gender, 2015 vote & 2016 vote.

DATA FOR ESTIMATING $p(t_i = 1|x_i)$

What proportion of each voter type will turn out to vote?

- EU 2016
  - 2015 British Election Study (BES) face to face survey is the only available high-quality probability sample with validated 2015 general election turnout.
- US 2016
  - 2012 Current Population Survey (CPS) is the largest available high-quality probability sample with 2012 turnout as a function of demographics.
- UK 2017
  - Pooled 2010 and 2015 BES.

Note: Lagged turnout needs to be a predictor in the turnout model to generate a reasonable mix of previous election voters/non-voters. We imputed lagged turnout using a mix of evidence and “informed priors”.

DATA FOR ESTIMATING $p(x_i)$

What proportion of the electorate is of each type?

- UK Data
  - 2011 UK Census 5% sample provides ~2m obs with joint distribution of *most* variables, Annual Population Survey (APS) about 200k/year.
  - Impute parliamentary constituency (and local authority for EU 2015), rake to known margins.
  - We impute 2015 turnout using 2015 BES face to face, 2015 vote using YouGov's 2015 polling data subject to constraint of actual election results in each constituency.
- US Data
  - 2010 US Census 1% sample provides ~2m obs with joint distribution of *most* variables.
  - We rake to update the marginal distributions using the 2015 American Community Survey (ACS).
  - We impute 2012 turnout and registration using current voter files, 2012 vote onto this using YouGov's 2012 polling data subject to constraint of actual election results in each congressional district.

POTENTIAL PROBLEMS

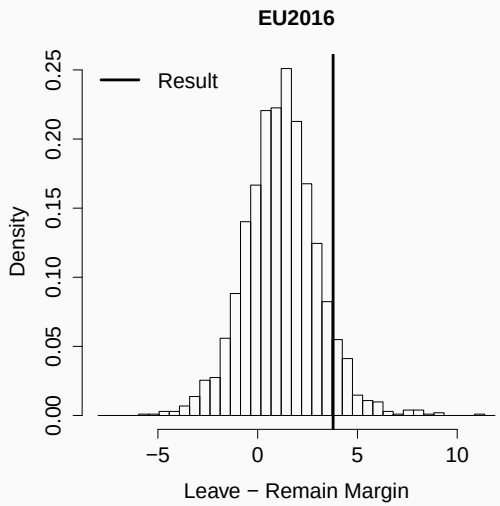
- Estimating $p(V_i|X_i, T_i = 1)$
  - YouGov panelists could have different preferences than the public, conditional on everything in the model.
  - Past vote measures for the panelists are crucial (but we have validated extensively, see: https://today.yougov.com/news/2016/11/01/beware-phantom-swings-why-dramatic-swings-in-the-p/)
- Estimating $p(T_i = 1|X_i)$
  - Turnout patterns will change from previous election
  - Turnout data from the last election may be unreliable
- Estimating $p(X_i)$
  - Many imputation/raking steps (with underlying assumptions)
  - Information about joint distributions of variables is partial
  - Aging the electorate is tricky

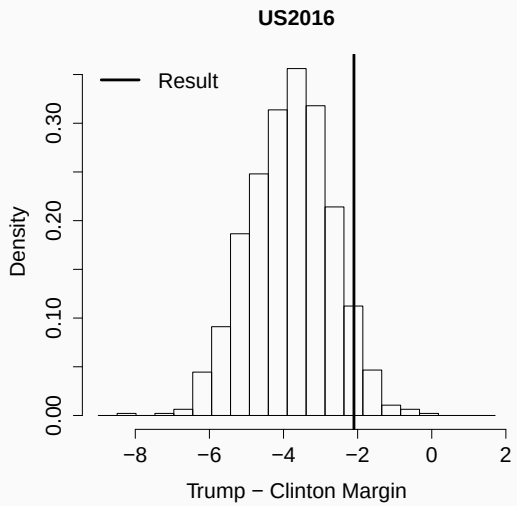
ALTERNATIVE WAYS TO APPLY MRP

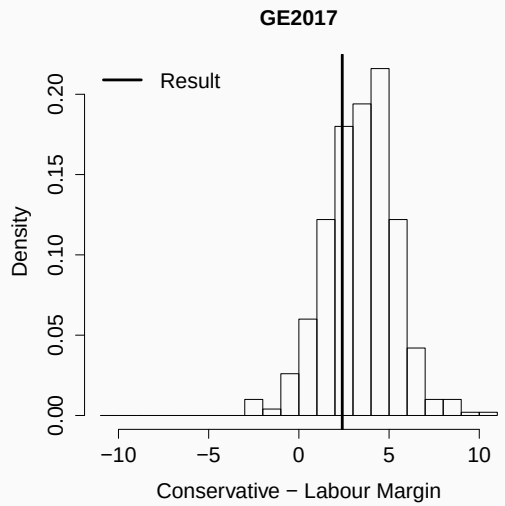
- Jointly model vote choice and not voting, post-stratification frame of eligible voters.
    - If survey is reasonably representative of voters, but not of non-voters, this will work less well than our approach.
    - If stated turnout intention is unreliable, this might work less well than our approach.
- Model vote choice, post-stratification frame of likely voters.
    - If the post-stratification frame is based on turnout in the last election, is effectively our approach, but difficult to use lag vote as a variable.

Results

NATIONAL VOTE SHARE MARGINS







	Vote Choice	Result	Estimate	Low	High
EU Referendum	Leave	51.9	50.6	48.8	52.4
	Remain	48.1	49.4	47.6	51.2
US Presidential	Trump	46.1	44.1	43.0	45.2
	Clinton	48.2	47.9	46.8	49.1
UK General	Conservative	43.4	41.6	39.2	43.9
	Labour	41.0	38.2	36.1	40.6
	Liberal Democrat	7.6	9.0	7.9	10.3
	UKIP	1.9	3.5	2.9	4.1
	Green	1.7	2.0	1.7	2.4
	SNP	3.1	3.8	3.4	4.2
	Plaid Cymru	0.5	0.5	0.4	0.6

Figure 2: National vote shares for major alternatives with mean posterior estimates and 95% predictive interval lower and upper bound.

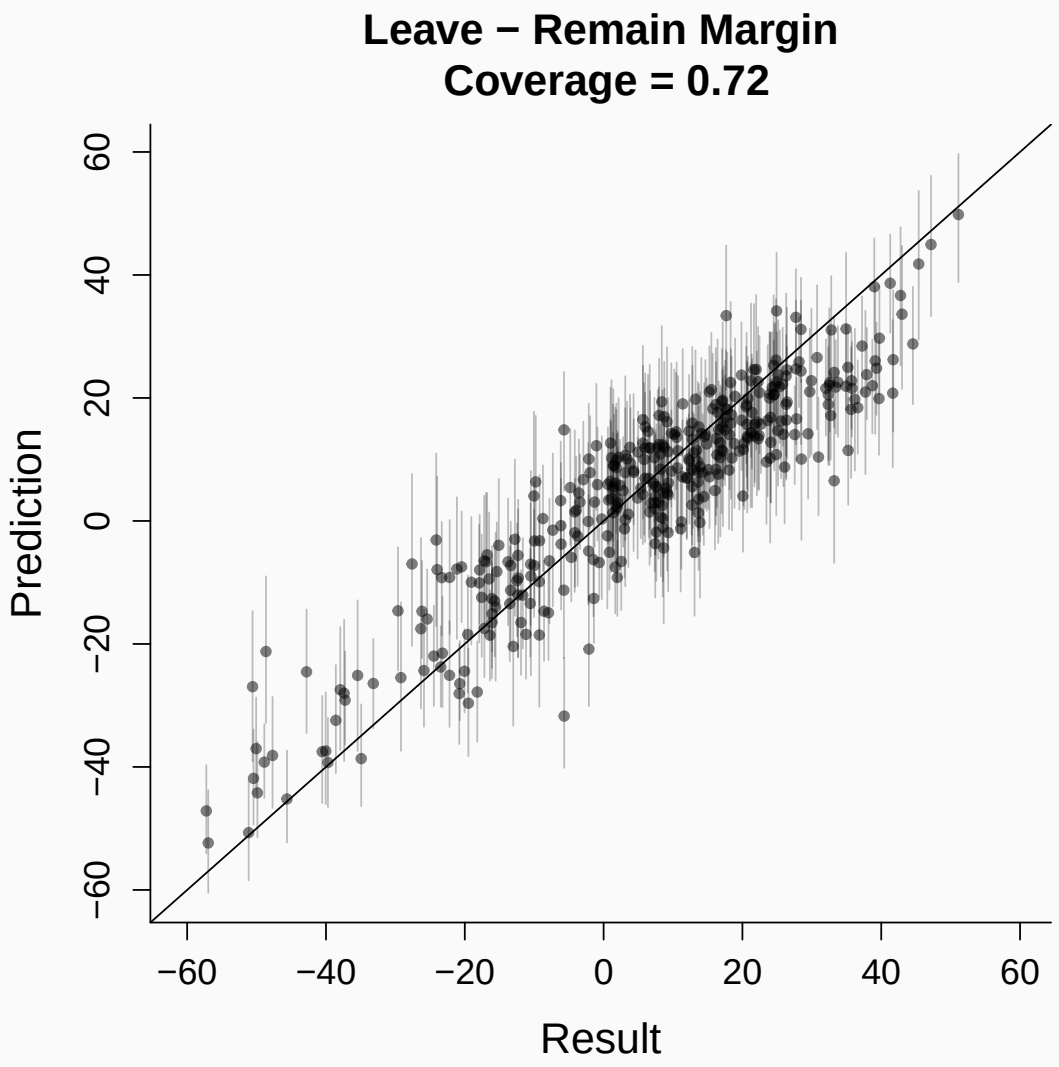


Figure 3: Local Authority Estimates versus Results for EU Referendum

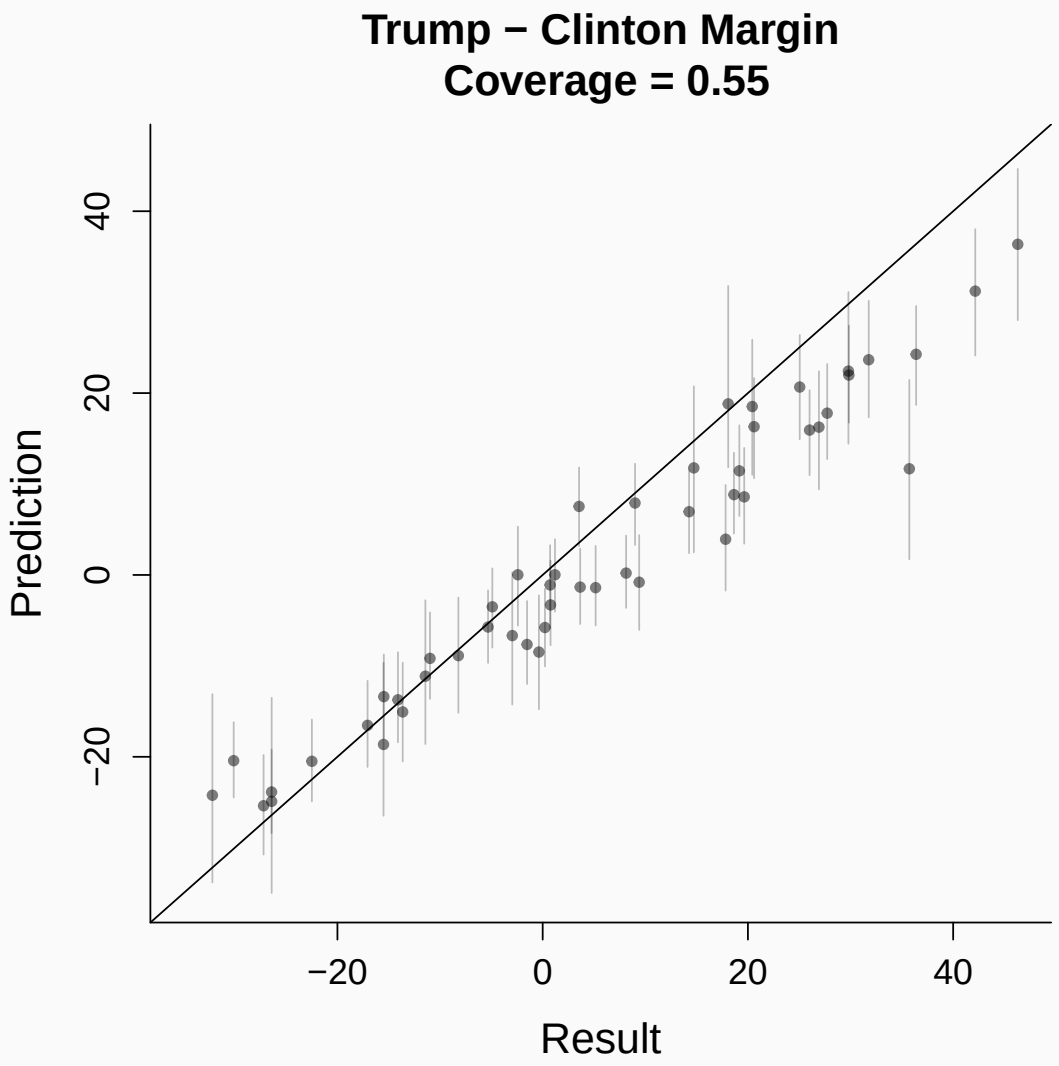


Figure 4: State Estimates versus Results for US Presidential Election

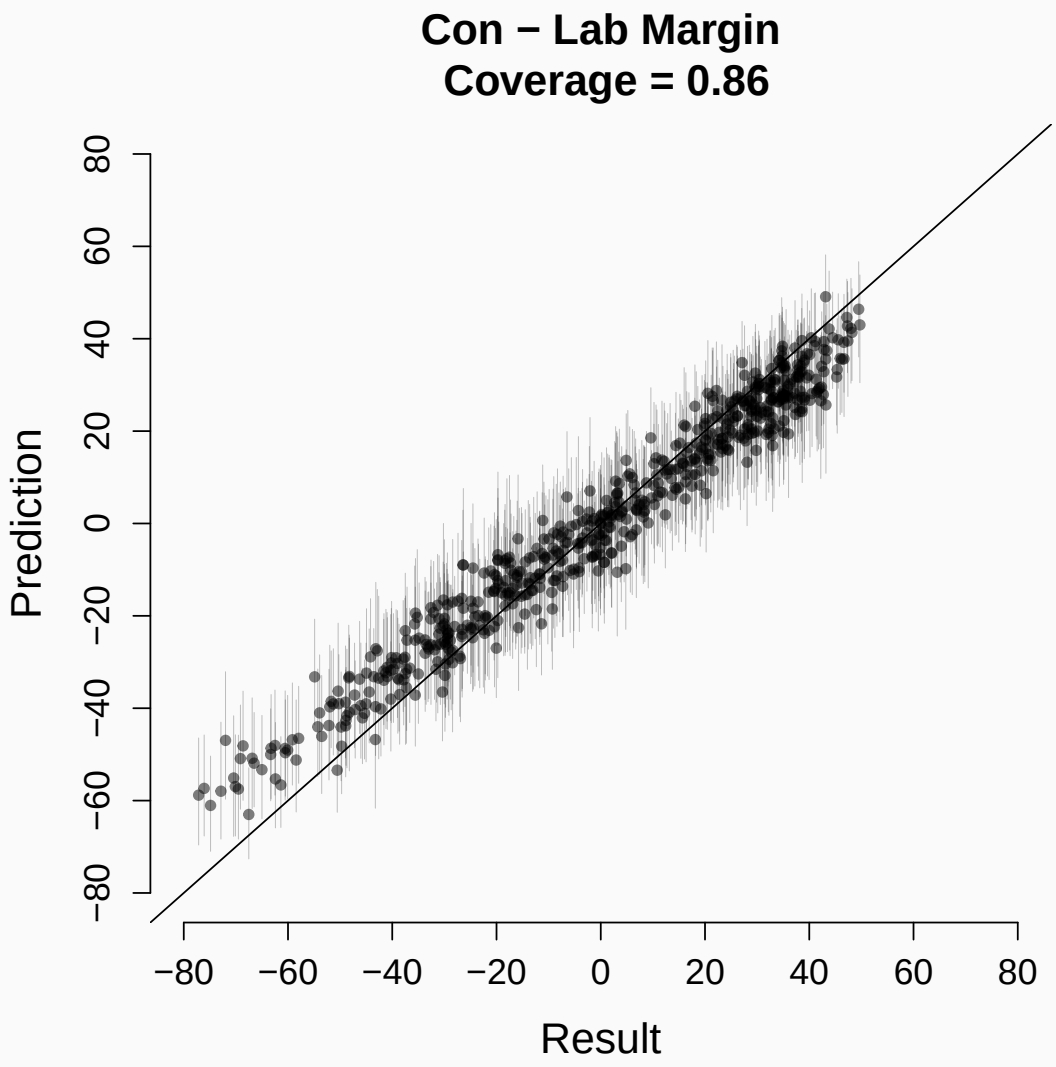


Figure 5: Constituency Estimates versus Results for UK General Election

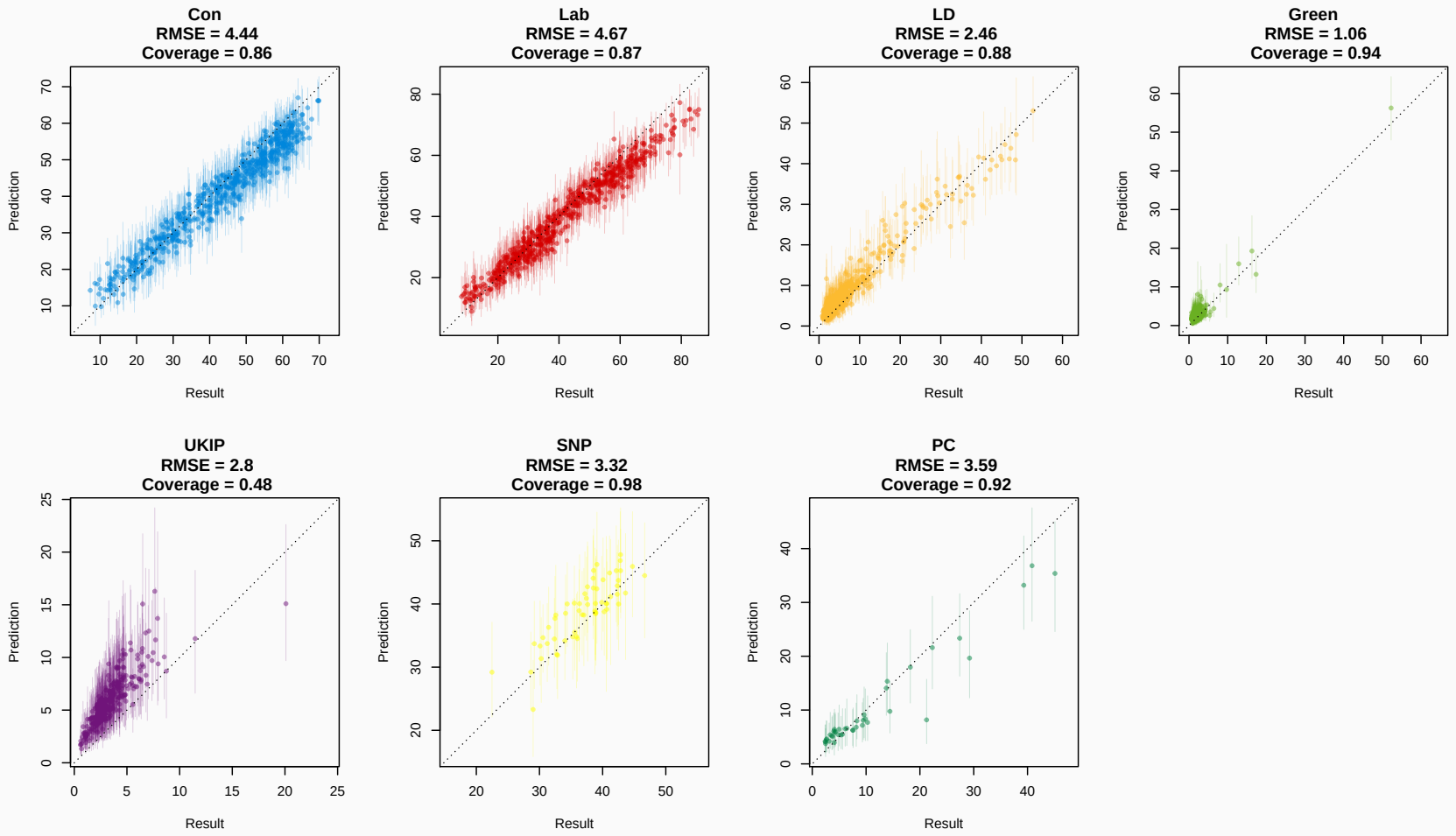


Figure 6: Constituency Estimates versus Results for UK General Election

Model	RMSE (Top Two Margin)
538 (polls plus)	7.0
Princeton Election Consortium	7.0
New York Times	7.0
538 (polls only)	7.1
YouGov	7.3
PollSavvy	8.0
HuffPost	10.7

Figure 7: US state level estimates - RMSE comparison

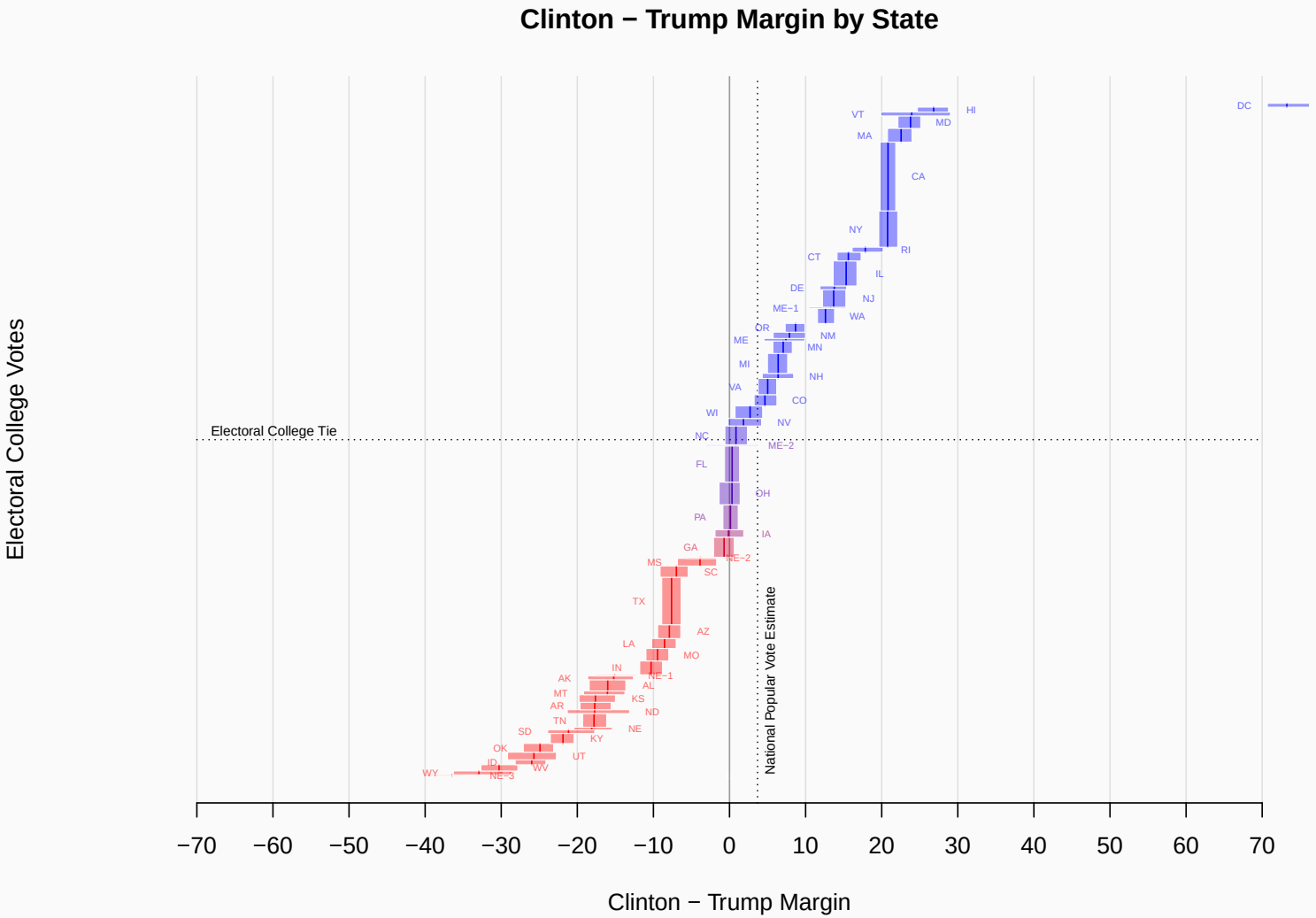


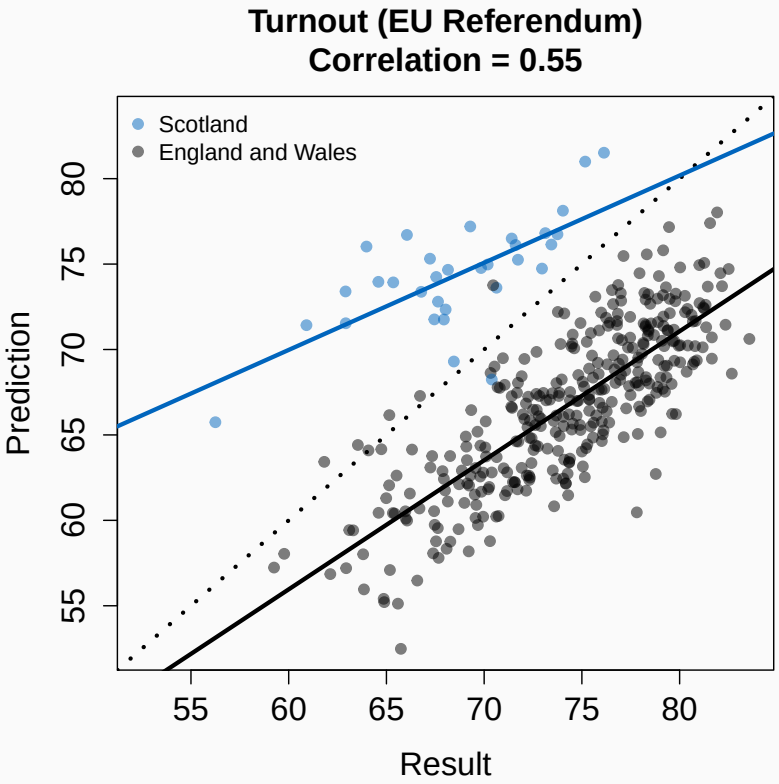
Figure 8: US state level estimates by Electoral Vote

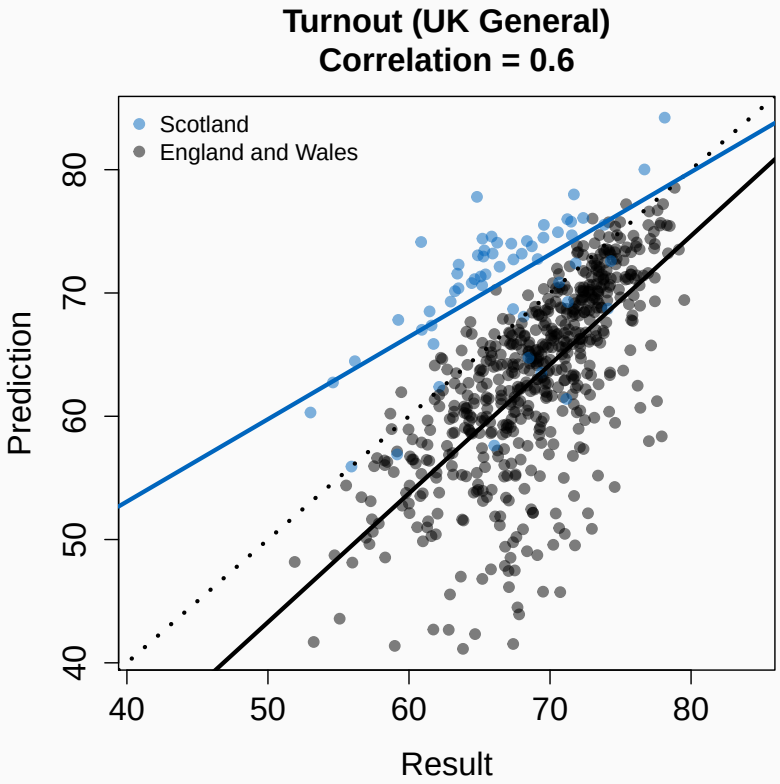
RMSE - UK

Model	Con	Lab	LD	UKIP	Green	SNP	PC	Other	Percent correct
YouGov model	4.4	4.7	2.5	2.4	0.9	3.3	3.6	1.6	92.9
Uniform swing (Regional)	4.6	4.1	3.6	3.8	1.9	3.8	2.9	1.9	91.6
Uniform swing (Country)	5.4	4.1	3.8	4.3	2	3.8	2.9	1.9	91.8
Uniform swing (GB)	5.9	4.7	3.8	3.8	1.8	11.9	3.3	1.9	91.1
Hanretty	5.3	6.1	3.7	1.9	1.9	4.7	4.1	2.3	86.2

Figure 9: UK constituency level estimates - RMSE comparison

TURNOUT


Turnout (EU Referendum)
Correlation = 0.55
Scotland
England and Wales
Prediction
Result
55
60
65
70
75
80


Turnout (UK General)
Correlation = 0.6
Scotland
England and Wales
Prediction
Result
40
50
60
70
80

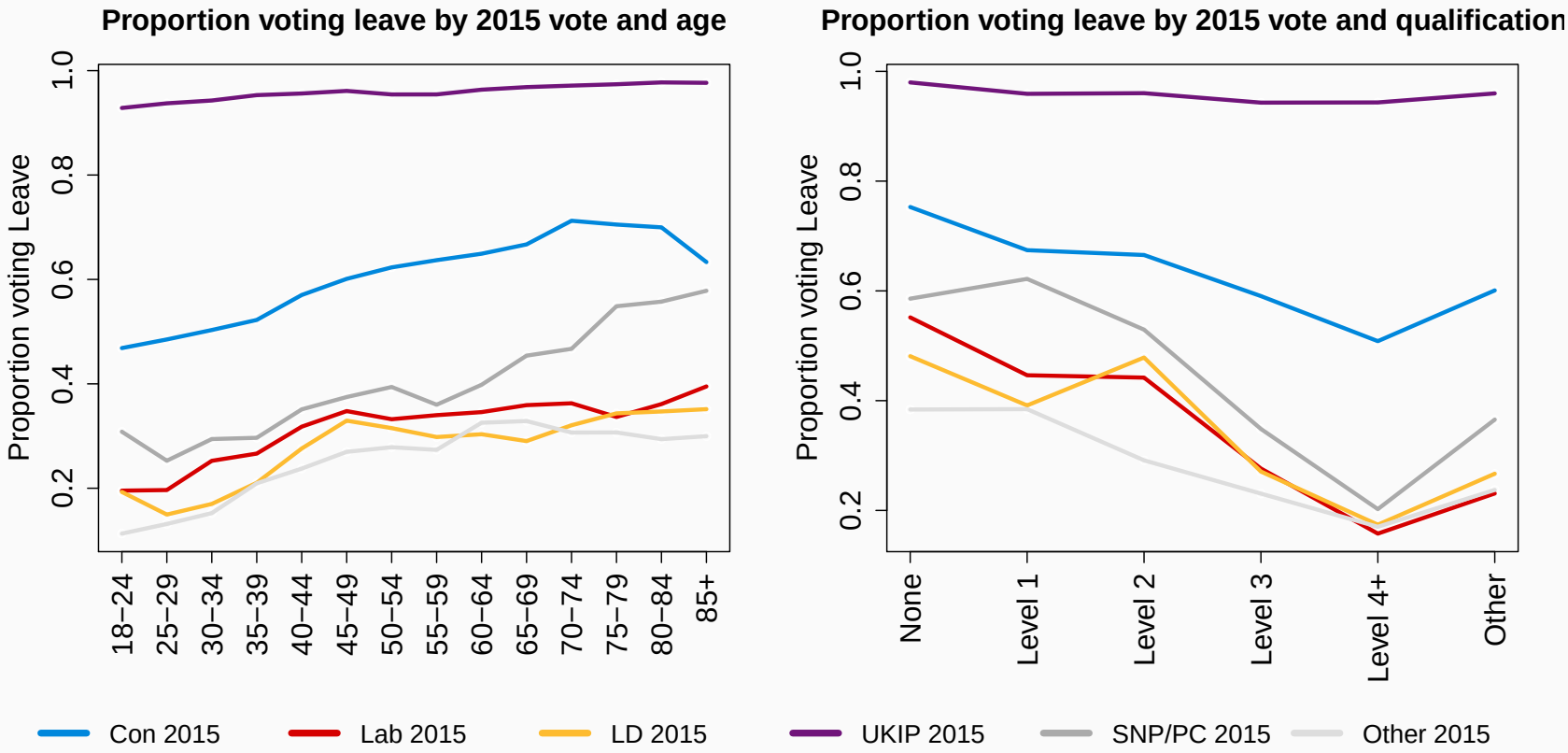


Figure 10: Leave Share by Age and Qualifications

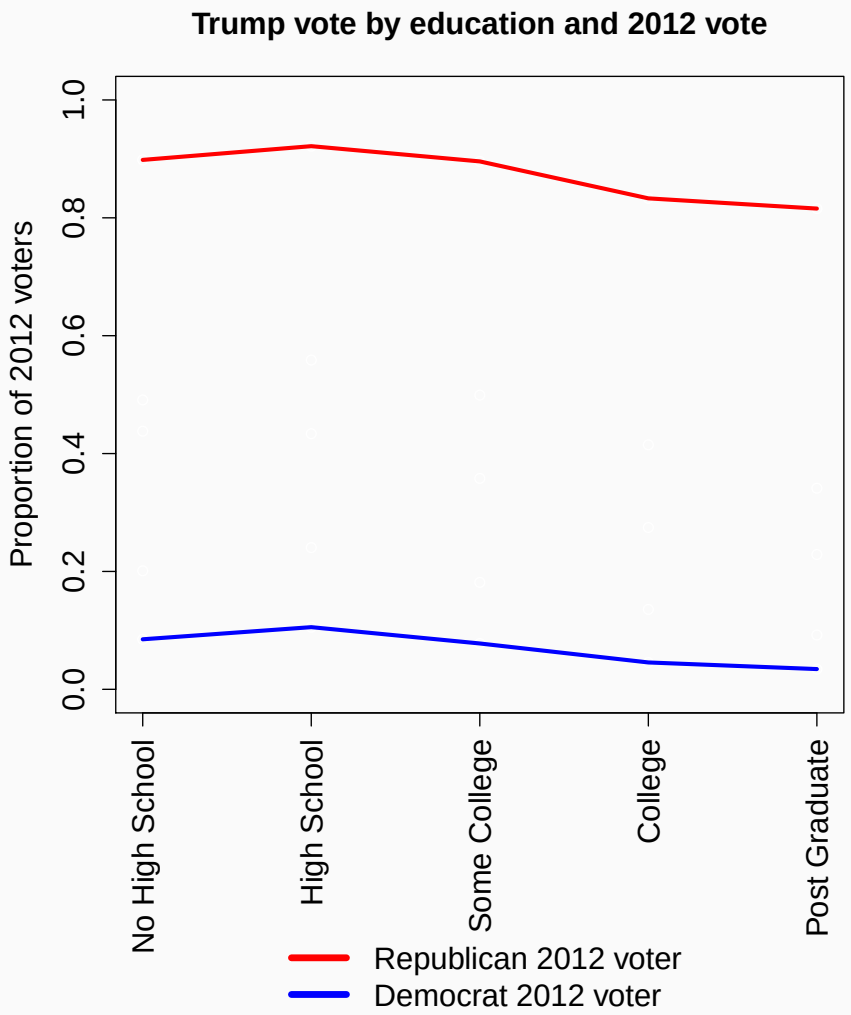


Figure 11: Trump Share by Education

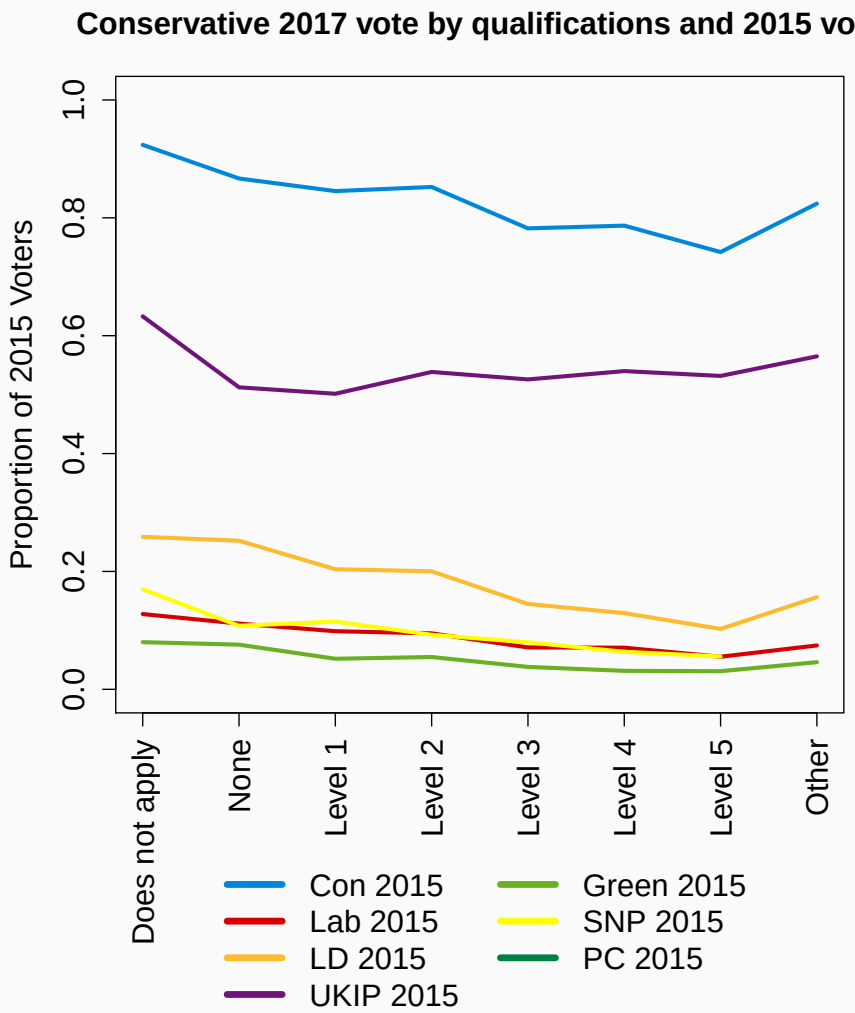


Figure 12: Con 2017 Share of Con 2015 voters by Qualifications

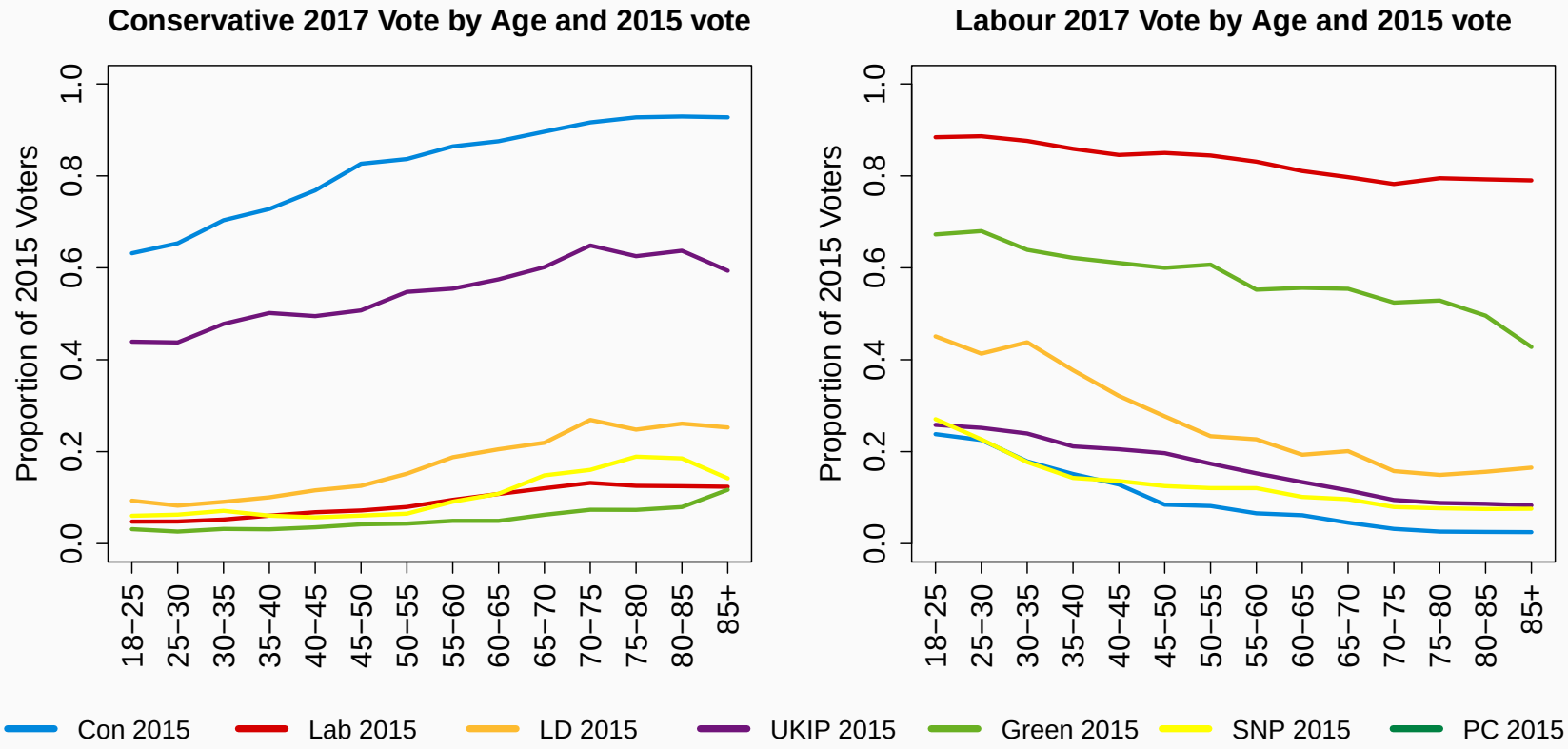


Figure 13: Con 2017 Share of Con 2015 voters by Age

CONCLUSIONS

- MRP worked reasonably well (if you had reasonable expectations!)
    - UK worked well at all levels of aggregation where we can validate, in both the referendum and the election.
    - US worked well at national level, but had more significant problems at state level.
    - We learned a lot about demographic patterns of voting.
- Our conservative solution to the turnout modelling problem was not perfect, but also was not a huge problem.
    - We need better signals of who is likely to vote *in high quality samples where we can properly calibrate them to the population*.

OBSERVATIONS ABOUT APPLICABILITY

- These methods for building national and sub-national estimates are most helpful when
    1. there are a large number of parallel FPTP elections occurring
    2. where electoral boundaries do not reflect enduring political communities
    3. where individuals are voting on the basis of national issues
- Which applications make sense?
    - UK general elections and US house elections are ideal applications
    - US presidential elections are ok, but moderate number of states and electoral college weighting make outcomes sensitive to relatively few sub-national units.
    - US senate and governor elections would be a struggle, candidates probably matter too much to make pooling across races useful.
    - Application to a national referenda or to PR systems may be useful for bias reduction reasons, but little payoff in terms of new estimands.

FUTURE DEVELOPMENT

1. Modelling turnout
    - We want to keep the baseline expectation that demographics of turnout do not change very much...
    - ...while incorporating information about which groups are stating more or less intention to turnout than last election.
2. Functional form
    - The binary/multinomial logit is not ideal, additive changes on that scale do not generate uniform swings at aggregate level.
    - Our vote choice models fit much better with interactions of state/constituency vote share at last election with individual-level vote choice at last election
    - We need to do more theoretical work on what is a good baseline functional form for these models
3. Uncertainty Calibration
    - This is sadly difficult to do in a principled way.

We will be back for US midterm elections in November 2018... and any UK general elections / referenda that might get called in the interim.