Revisiting the UK EU Membership Referendum (Brexit) Poll Tracker

Michaelino Mervisiano

Abstract

On the 23rd June 2016, the United Kingdom (UK) European Union (EU) membership referendum resulted in 51.9% of voters voted to leave EU—popularly termed as Brexit. Given its significant implications, correctly predicting Brexit was crucial but most pollsters predicted incorrectly. This paper assesses whether Brexit was evident and predictable from the pre-referendum polls data. Unlike previous studies—whose analytical tools are limited to latest poll analysis, descriptive statistics, point estimate, and simple linear regression—this project use more robust and sophisticated statistical methodologies

Statistics with Data Science Consultancy Project

Michaelino Mervisiano

Supervised by

Dr. Miguel de Carvalho

Master of Science
Department of Mathematics and Statistics
University of Edinburgh

2017

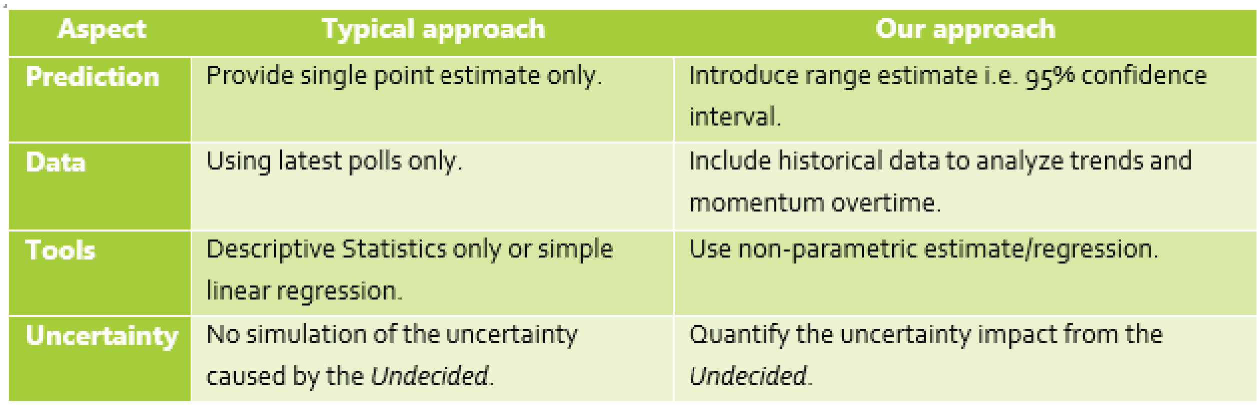
Executive Summary

1. Importance of Predicting Brexit Correctly

On the 23rd June 2016, the United Kingdom (UK) European Union (EU) membership referendum resulted in 51.9% of voters voted to leave EU—popularly termed as Brexit. Given its significant implications, correctly predicting Brexit was crucial but most pollsters predicted incorrectly.

This paper **assesses whether Brexit was evident** and predictable from the pre-referendum polls data. Unlike previous studies—whose analytical tools are limited to latest poll analysis, descriptive statistics, point estimate, and simple linear regression—this project use more robust and sophisticated statistical methodologies as shown by Table1.

Table1*: Comparison of Methodology*



Aspect	Typical approach	Our approach
Prediction	Provide single point estimate only.	Introduce range estimate i.e. 95% confidence interval.
Data	Using latest polls only.	Include historical data to analyze trends and momentum overtime.
Tools	Descriptive Statistics only or simple linear regression.	Use non-parametric estimate/regression.
Uncertainty	No simulation of the uncertainty caused by the *Undecided*.	Quantify the uncertainty impact from the *Undecided*.

2. Analysing Pre-referendum Polls Data

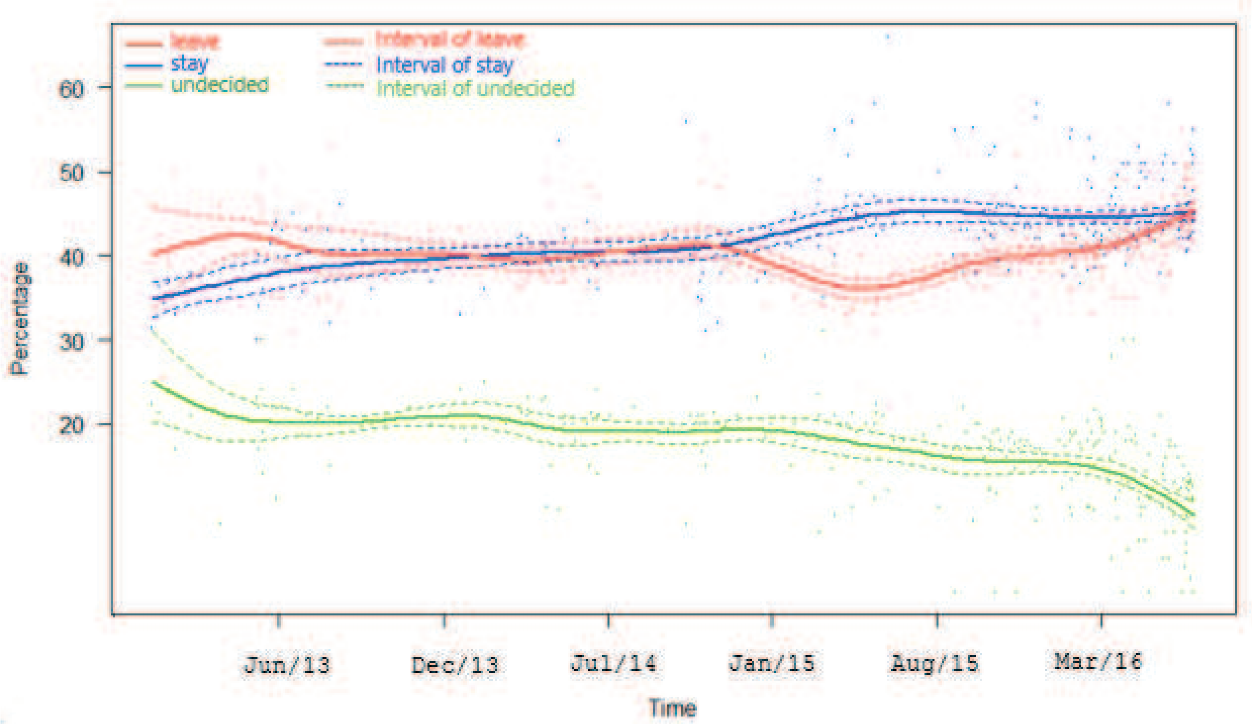
Dataset for this study are 261 pre-referendum polls from 15 pollsters between January 2013 and June 2016. From data analysis, we observe:

1. ***Latest Poll***
   Consider only the most recent polls held on 22nd Jun 2016. From the overlapping CI's, we clearly see there is a strong possibility of Brexit.
2. ***Combined Poll***
   Combine all polls from the oldest to the newest and treat it as single sample, weighted by sample size. While the CI's are not overlapping, we see significant proportion of the *Undecided* (16.48%) as a big source of uncertainty.
3. ***Weighted Poll***
   Combine all polls from the oldest ones to the newest ones but consider the evolving trend of the leave/stay decision. Our approach is applying nonparametric regression. Using Nadaraya-Watson Estimator and Local Polynomial Regression to get the best fit estimation.

Figure1 shows the trends of *Stay*, *Leave*, and *Undecided* over time. We see overtime, the number of *Undecided* is steadily declining while *Leave* growing fast in last few months and slightly over *Stay* as Time approach to voting day.

Next, we interest to make a prediction of referendum outcome with nonparametric regression.

Figure1: *Polling Trend Overtime Using Local Polynomial Regression*



3. Predicting Brexit from pre-referendum polls

We use local polynomial regression to predict the Brexit result. Our analysis considers three different scenarios of treating the *Undecided*.

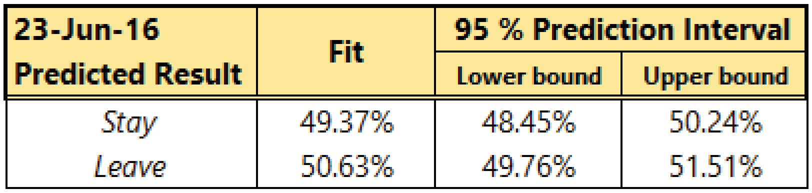
1. **No splitting**

   Assume the *Undecided* will not vote. The model correctly predicts *Leave*, however the *Stay* and *Leave* CI's still overlapping.

2. **50:50 Splitting**

   Assume 50% of *Undecided* will go to *Leave* and 50% to *Stay* (maximum uncertainty). The model correctly predicts *Leave*, and the CI's overlap become narrower. Table2 shows that point prediction clearly suggest that UK public prefer to leave EU. Based on, prediction interval, we see that there are overlap between *Stay* and *Leave*. It means that Brexit is not a surprising result. The historical polls did suggest it might happen.

Table2: *Prediction Result based on 50:50 Splitting*



23-Jun-16 Predicted Result	Fit	95 % Prediction Interval	
		Lower bound	Upper bound
Stay	49.37%	48.45%	50.24%
Leave	50.63%	49.76%	51.51%

4. Conclusion and Recommendation

As conclusion:

1. Given the overlapping confidence intervals between *Leave* and *Stay,* Brexit is actually not a surprising result.
2. Brexit is actually predictable from the historical polls data if one uses more robust methodologies. Many pollsters made incorrect prediction because they rely on very basic methodology.

Going forward, to ensure better predictive accuracy, we suggest pollsters to:

- Use confidence interval estimates.
- Not ignore the *Undecided* when its proportion is big.
- Analyse trends overtime, not relying on latest poll only.
- Use non-parametric estimation/regression where appropriate.

Table of Contents

1. Introduction

This section describes the background of the study, why it is important, the objectives of the study, as well as summary of the research questions will be answered.

1.1. Overview

There is shocking result of the EU membership referendum by UK. On the 23rd June 2016, the United Kingdom of Great Britain and Ireland (UK) held a referendum on the UK's European Union (EU) membership—which is popularly known as the Brexit referendum. Brexit is an informal but popular term for potential withdrawal UK from the EU. The referendum resulted in 51.9% of voters voted to leave EU. Following referendum result, government of UK initiated the official EU withdrawal process on 29th March 2017, which put UK on course to complete the withdrawal process by 30th March 2019.

The result of Brexit referendum was considered shocking for three main reasons. First, the unilateral decision to withdraw from EU is in contrast with UK's reputation as an open-minded, tolerant and outward-looking country.

Second, referendum result has significant impacts. For example, it created significant volatility in the financial markets. On 24th June 2017, the British pound sterling dropped from $1.50 to $1.37 – the biggest move for the currency in any two-hour period in history. Similarly, investors in worldwide stock markets lost more than the equivalent of US$ 2trillion on 24th June 2016, making it the worst single-day loss in history, in absolute term. Furthermore, it created a political crisis. Soon after the result, David Cameron, then the Prime Minister for UK, announced his resignation, having unsuccessfully campaigned for a "Remain" vote. The unwillingness of Scotland and Northern Ireland to leave the EU may lead to a constitutional crisis for UK.

Third, the result is in contrast with many predictions that had been issued by various major pollsters. Given its significant impacts, ability to correctly predict the referendum result was important. Yet, most major pollsters—such as YouGov, Populus, ComRes, ORB, and Survation—failed to predict the outcome correctly and only two major pollsters, TNS and Opinium, correctly predicted the outcome. As the result, many stakeholders have raised questions about the reliability and validity of the polls. Brexit also has become a press release spell disaster for the commercial pollsters industry.

1.2. Objective of This Project

Given the importance of predicting correctly the result from polling data for occasion such as presidential/prime minister election, important referendum, or popular votes in the future, it is important for us to understand whether the Brexit referendum result can be predicted based on polls. This paper assess whether the result of the Brexit referendum was evident from the historical polls. The main goal of this project is to address the following questions:

1. Was the Brexit referendum result obvious from the historical poll data?
2. Can the Brexit referendum result be predicted based on the historical poll data?
3. What likely went wrong in the Brexit referendum result prediction?
4. How much can we trust the polls?
5. What can be done differently in the future, to avoid such incorrect prediction from most major pollsters?

In addition, to meet the clients' need on estimation and prediction, we are producing a model which will smoothed data in a way that estimates add up to one (100%). The sum of each estimation from respondent's answer (*Stay*, *Leave*, *Undecided*)[1] will be precisely equal to one.

1.3. What This Project Will Do Differently

There are only few studies revisited the Brexit poll data to determine whether the result of Brexit referendum is predictable. Please refer to Section 2 for further details.

In general, previous studies rely on very basic statistical methodologies. The analytical tools employed are often limited to latest poll analysis, descriptive statistics, point estimate, and simple linear regression.

We believe more robust and sophisticated statistical methodologies will significantly improve the predictive accuracy. Improvements we introduce in this project are:

1. Use range estimate, in addition to point estimate.
2. Assimilate historical data (beyond the latest poll).
3. Employ non-parametric regression to predict the result.
4. Consider uncertainties resulted from the *Undecided* respondents.
5. Analyse trends and momentum overtime.

In the end, this project will assess whether Brexit was predictable based on historical poll data if more robust and sophisticated methodologies have been employed.

[1] Please refer to Section 3.1 for further details.

2. Background and Literature Review

This section discusses literatures related to analyzing on polling data and predicting the result of Brexit referendum.

2.1. Literature Review

There is enduring academic interest in the problem of understanding and predicting a voting outcome. For long time, authors have used regression models on nationwide polling data to forecast the outcome of popular vote [CW90]. However, forecasting with linear regression may not feasible on every cases when the distribution of data is not parametric [C96]. Therefore, there is a need to use an approach would fit condition where parametric could not be held. One of prominent method is fitting the polls data with smooth curves which could be computed with kernel estimation [BH14].

In fact, studies claimed that although polling data are inevitably flawed, they can still provide much insight on about the national and regional trends—as long as one is aware of this and pay attention to the associated validity of statistical inference [FS16]. Another study emphasised on the importance of confidence intervals and uncertainty quantification from any prediction made on polling data [FLB16].

As polls data before voting days are not updated on a consistent basis, it raises a concern about how to incorporate the new and old data? Should we use only the latest poll? Should we use all the historical polls?

An intuitive approach is proposed three different data assimilation methods [CW08]:

1. **Latest Poll:** Examine just the latest poll.
2. **Combined Poll:** Consolidate every single past polls, but regard it as a single sample. The re-weighting is solely from sample size. There is no adjustment affect from time period.
3. **Weighted Poll:** Consolidate every single past polls, but the re-weighting is adjusted based on time period, depending on the day the poll is taken. Observed the poll based on trend of time period.

2.2. Previous Studies on Efforts to Predict/Revisit the Brexit Result

Since the referendum result carries significant impact, several studies have tried to cover the topic on predicting the Brexit referendum result.

In March 2016, Simon Jenkins claimed that the voters were acting based on gut instinct alone. As the implication, it is become impossible to predict the result of the referendum. He used descriptive analysis only [JS16].

In April 2016, Rohn Jonston et al analysed the YouGov polling data between 2015 and Q1 2016 using cross-tabulation and concluded that the young and the better educated voters appear much likely to vote for Stay, while the elderly and the less educated voters are more likely to vote for Leave. However, they did not provide any prediction on the actual result. In their analysis, they used descriptive statistics only [JRA16].

HD Clarke and M Goodwin also pointed out that while the YouGov poll data they used in April 2016 showed a clear majority support for Stay, by the referendum date, the difference has shrunk significantly. The authors only described this fact and did not conduct any inferential statistic tests on the trend [CG16].

In August 2016, Joh Fry claimed that the Brexit referendum was supposed to be a simpler problem than general election problem, i.e. no complications of multi-party system and individual constituency-level effects, but is still challenging to predict the result. He used linear regression in his analysis, he predicted the result will be 48.7% for Leave and 51.3% for Stay. He then proceed to claim that the Brexit referendum is not predictable as the (linear-regression) prediction would ultimately turn out to be wrong [F16].

Compared to other studies, John Fry had moved beyond descriptive statistics and point estimate. He considered prediction and historical trends over time, not only latest poll. However, we observed two drawbacks with John Fry's analysis: 1) he employed linear model when the trend is not necessarily linear; 2) he ignored Undecided respondents – which is a big source of uncertainty.

3. Analysis of the Polling Data

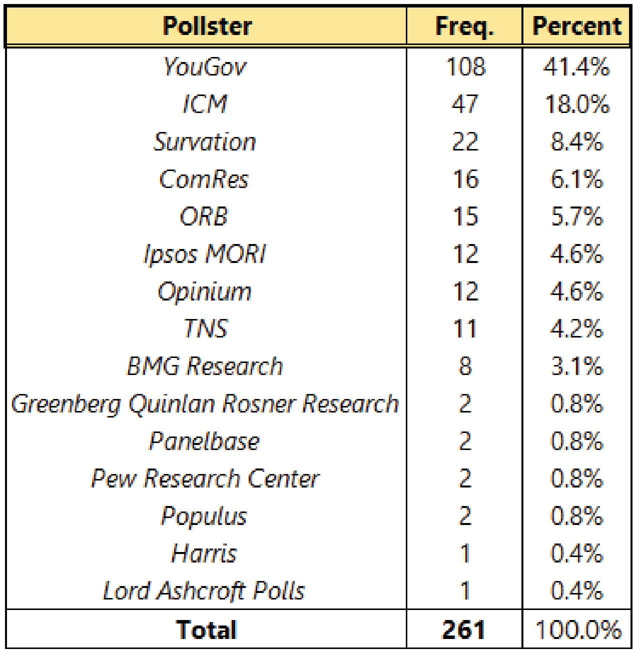
This section reviews the pre-referendum polling data and analyses whether the result of Brexit is evident from it.

3.1. Data Description

Our pre-referendum polling dataset is composited from 15 pollsters. The dataset contains 267 historical polls results for UK citizens between 9th September 2010 and 22nd June 2016. In the polling, the respondents were asked about their opinion on staying vs. leaving the European Union (EU).

The polling became more frequent since January-2013. Thus, to eliminate data sparsity issue, we only used data from January-2013 to June-2016 (**261 observations**). Table3.1.1 shows frequency of polls each pollster held during 2013-2016. YouGov did the most surveyed from 2013-2016 (with 108 polls), 41.4% of our data come from YouGov. In second and third place, there are ICM and Survation who did 47 and 22 surveyed respectively.

Table3.1.1: *Frequency of Polls from Various Pollsters*



Pollster	Freq.	Percent
YouGov	108	41.4%
ICM	47	18.0%
Survation	22	8.4%
ComRes	16	6.1%
ORB	15	5.7%
Ipsos MORI	12	4.6%
Opinium	12	4.6%
TNS	11	4.2%
BMG Research	8	3.1%
Greenberg Quinlan Rosner Research	2	0.8%
Panelbase	2	0.8%
Pew Research Center	2	0.8%
Populus	2	0.8%
Harris	1	0.4%
Lord Ashcroft Polls	1	0.4%
Total	**261**	100.0%

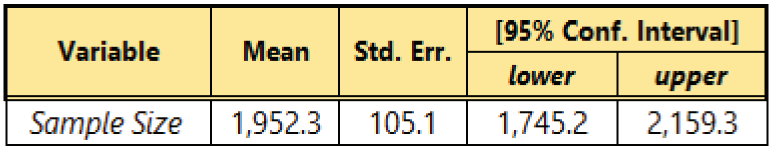
The pollsters mostly used internet and telephone survey to carry out the polling. The dataset contains six variables, *Stay, Leave, Undecided, Date, Pollster,* and *Sample size*. The description are as follow:

1. **Stay:** Percentage of respondents who believe UK should stay in EU.

2. **Leave:** Percentage of respondents who believe UK should leave EU. If *Leave* is greater than *Stay*, it means BREXIT would happen.
3. **Undecided:** Percentage of respondents who has undecided.
4. **Date:** Date when poll was held.
5. **Pollster**: Company who held the poll.
6. **Sample size:** Number of respondents.

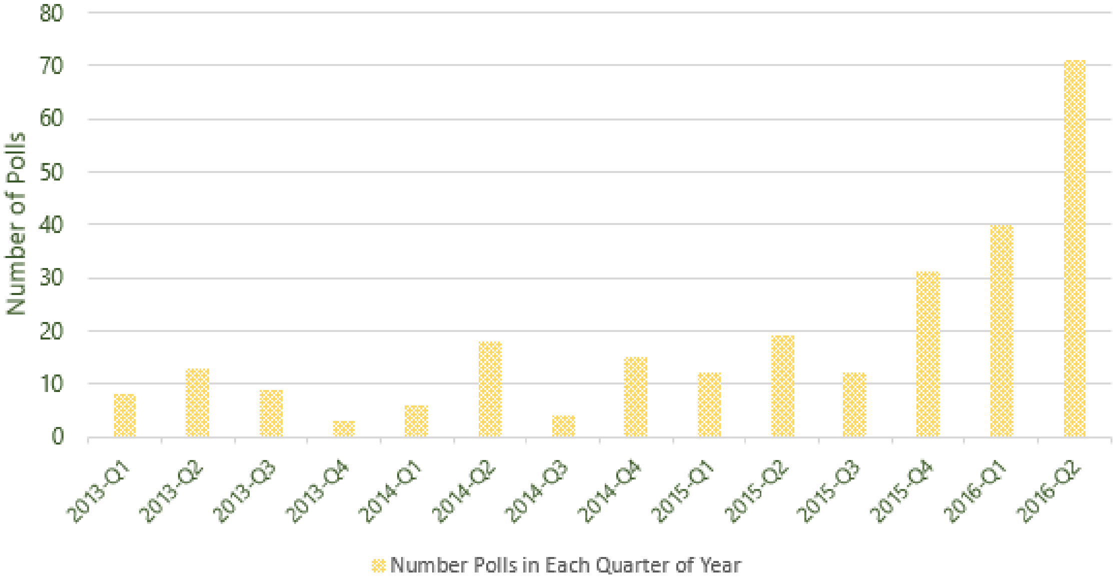
Table3.1.2 shows the mean of sample size is 1,952.3 with range in [1,745.2, 2,159.3]. As date closer to voting date, number of polls is increasing as shown in Figure3.1.1.

Table3.1.2: *Sample Size: Mean & Confidence Interval*



Variable	Mean	Std. Err.	[95% Conf. Interval]	
			lower	*upper*
Sample Size	1,952.3	105.1	1,745.2	2,159.3

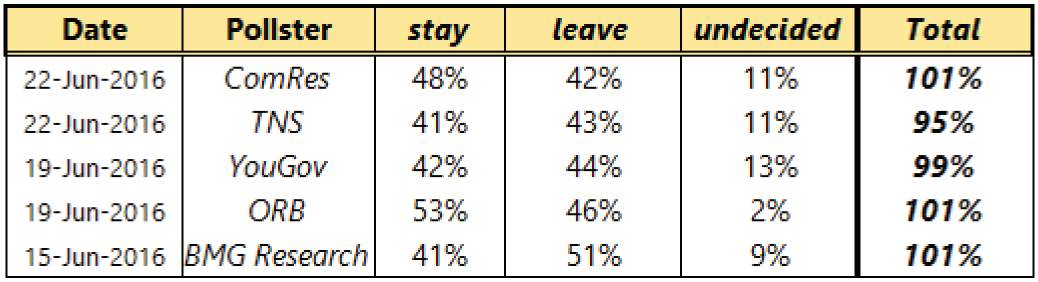
Figure3.1.1: *Number of Polls in Each Quarter of Year*



3.2. Data Adjustment: Projection of a Point on a Plane

Unfortunately, the polling data contain slight mistakes. In 76 polling, sum of *Stay, Leave,* and *Undecided* is not equal to 100%. A few examples of observations who suffered of this issue are shown in Table3.2.1. This issue need to be solved since one of our clients' need is producing a model which will smoothed data in a way that estimates **add up to one**. Therefore, we will adjust the *Stay*, *Leave*, *Undecided* such that their sum equal to 100% using a point on a plane projection technique.

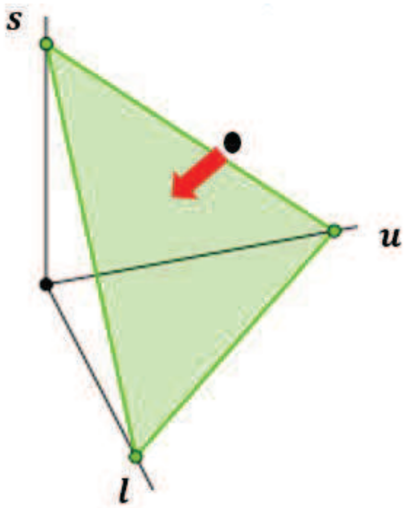
Table3.2.1: *Example of Observations which sum of Stay, Leave, and Undecided is not equal to 1*



Date	Pollster	*stay*	*leave*	*undecided*	*Total*
22-Jun-2016	*ComRes*	48%	42%	11%	***101%***
22-Jun-2016	*TNS*	41%	43%	11%	***95%***
19-Jun-2016	*YouGov*	42%	44%	13%	***99%***
19-Jun-2016	*ORB*	53%	46%	2%	***101%***
15-Jun-2016	*BMG Research*	41%	51%	9%	***101%***

Consider s, l, u denoted as variables *Stay*, *Leave*, *Undecided*. We want to project the points into plane $s + l + u - 1 = 0$. We wish to get projection such that $\vec{s} + \vec{l} + \vec{u} = 1$. Figure3.2.1 shows illustration of point that will be projected to the plane.

Figure3.2.1: *Illustration of Projecting Point to Plane (Simplex)*



Further detail of projection concept can be found in Appendix 1. Now, **as an example**, we will work on **second observation** that shown in Table3.2.1. It showed that $s + l + u = 0.95$, thus we need to compute normalization of point to plane, $t_0 = -\frac{0.41+0.43+0.11-1}{1^2+1^2+1^2} \approx 0.017$. Then, we can compute the projection result for each variables, as follow:

$$\vec{s} = 0.41 + t_0 = 0.4266667$$

$$\vec{l} = 0.43 + t_0 = 0.4466667$$

$$\vec{u} = 0.11 + t_0 = 0.1266667$$

Finally, as a result, we got $\vec{s} + \vec{l} + \vec{u} = 1$. With this projection, now the sum of *Stay, Leave,* and *Undecided* will equal to 100%. We did this projection to **all observations** and use the projection **results for all next analysis**.

3.3. General Assumptions on the Polling Data

We would like to clarify the general assumptions before we start the analysis in next section:

1. The pollsters used relatively similar sampling methodology. Therefore, we can reasonably combine the polls.
2. The pollsters followed robust sampling technique to ensure sampling randomness and representativeness. Therefore, each polling result is a reasonable representative of the voters.
3. The polling results between different pollsters are independent and identically distributed. In other way, a polling result from one pollster will not influence the results for other pollsters.
4. The polling results are not auto-correlated i.e. the result of previous polls are not affecting the result of later polls.

We will discuss more about the validity of these assumptions in Section 5.

3.4. Analysis of the Pre-referendum Polling Result

In Section 2, we have discussed common approach to working on polling data. To comprehend as our pre-referendum polls were continually being updated for months prior to the real referendum, we must consider a way to assimilate the newer and older information. In this analysis, we will follow the proposal suggested by Christensen & Florence (2008) by considering three methods: ***Latest Poll***, ***Combined Polls,*** and ***Weighted Poll***.

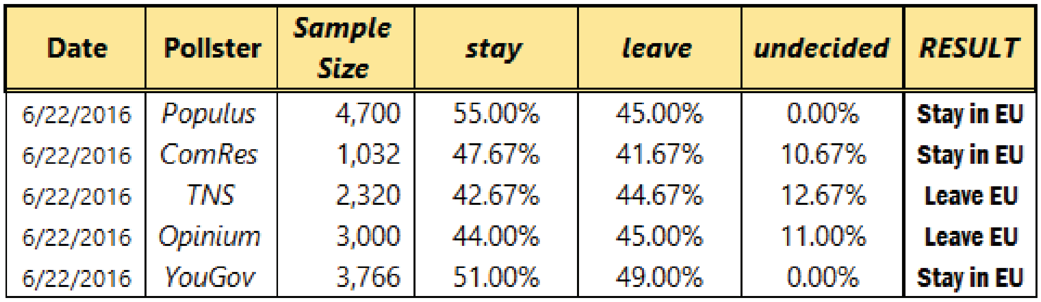
3.4.1. Latest Poll

Here, we assume only the latest poll matters i.e. it replaces all previous polls. In this case, the latest polls are five polls conducted on 22nd June 2016, a day before the Brexit referendum. Table3.4.1 shows the result of these polls. As shown, the result is inconclusive. Three pollsters predicted "Stay" and two pollsters predicted "Leave" as the result.

However, we believe point estimation is insufficient to infer the result. Thus, we compute the confidence interval for **decided results** only with binomial proportion. Table3.4.2 provides the rescaled percentage of decided respondents (rescaling was done in considering only *Stay* and *Leave* voters based on sample size).

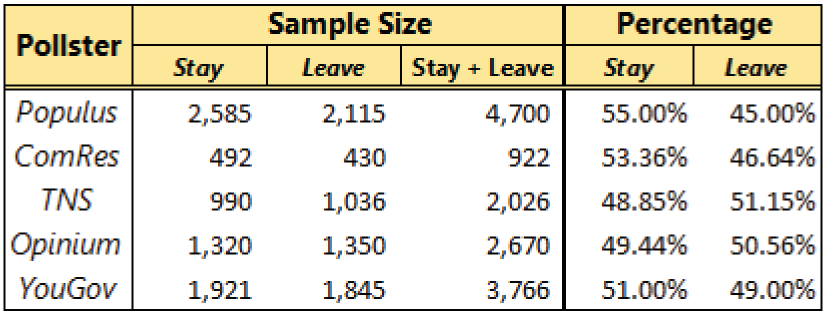
We consider three different methods: Wald, Agresti-Coull, and Clopper-Pearson confidence interval (CI). Details on equation and formula on calculating 95% confidence interval can be found in Appendix 2.

Table3.4.1: *Result of Latest Poll – Point Estimate*



Date	Pollster	Sample Size	stay	leave	undecided	RESULT
6/22/2016	Populus	4,700	55.00%	45.00%	0.00%	Stay in EU
6/22/2016	ComRes	1,032	47.67%	41.67%	10.67%	Stay in EU
6/22/2016	TNS	2,320	42.67%	44.67%	12.67%	Leave EU
6/22/2016	Opinium	3,000	44.00%	45.00%	11.00%	Leave EU
6/22/2016	YouGov	3,766	51.00%	49.00%	0.00%	Stay in EU

Table3.4.2: *Latest Poll – Point Estimate*
Decided Voters Only

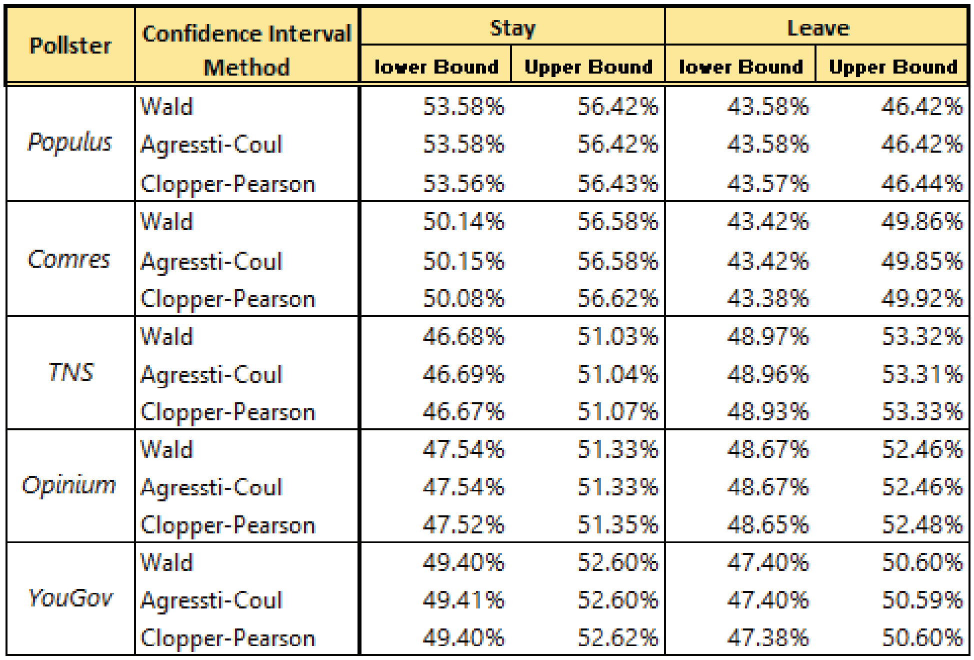


Pollster	Sample Size			Percentage	
	Stay	Leave	Stay + Leave	Stay	Leave
Populus	2,585	2,115	4,700	55.00%	45.00%
ComRes	492	430	922	53.36%	46.64%
TNS	990	1,036	2,026	48.85%	51.15%
Opinium	1,320	1,350	2,670	49.44%	50.56%
YouGov	1,921	1,845	3,766	51.00%	49.00%

Table3.4.3 shows the result of confidence intervals from three methods for 5 pollsters. Based on our data, we can observe that outcomes of Wald and Agresti-Croull are really similar to each other. Outcome from Clopper-Pearson is slightly different, but it's too small (not significant difference). Thus, we conclude all three methods gives very similar result. Next, to show how much overlapping between interval of *Stay* and *Leave*, we will use box plot as shown in Figure3.4.1. This box plot was created based on result from Clopper-Pearson method. From Figure3.4.1, we see that, for all pollsters except Populus, *Stay* and *Leave* CI significantly overlapped. This indicates it is very difficult to predict the result based on latest poll alone. Therefore, we start to include the historical polls in next analysis.

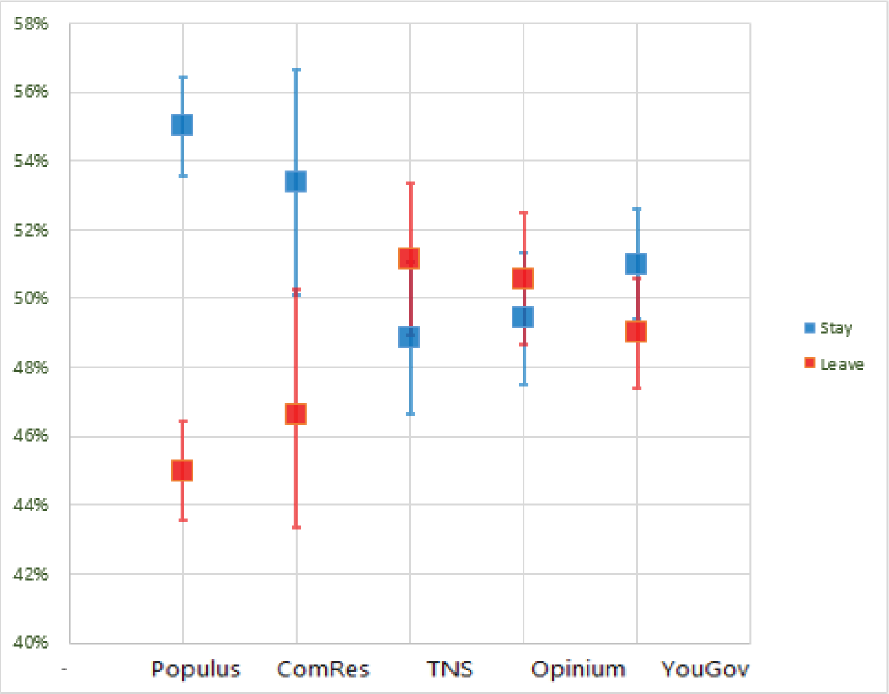
Table3.4.3: *Latest Poll Result-Confidence Interval*

Wald, Agressti-Coul, Clopper-Pearson



Pollster	Confidence Interval Method	Stay		Leave	
		lower Bound	Upper Bound	lower Bound	Upper Bound
Populus	Wald	53.58%	56.42%	43.58%	46.42%
	Agressti-Coul	53.58%	56.42%	43.58%	46.42%
	Clopper-Pearson	53.56%	56.43%	43.57%	46.44%
Comres	Wald	50.14%	56.58%	43.42%	49.86%
	Agressti-Coul	50.15%	56.58%	43.42%	49.85%
	Clopper-Pearson	50.08%	56.62%	43.38%	49.92%
TNS	Wald	46.68%	51.03%	48.97%	53.32%
	Agressti-Coul	46.69%	51.04%	48.96%	53.31%
	Clopper-Pearson	46.67%	51.07%	48.93%	53.33%
Opinium	Wald	47.54%	51.33%	48.67%	52.46%
	Agressti-Coul	47.54%	51.33%	48.67%	52.46%
	Clopper-Pearson	47.52%	51.35%	48.65%	52.48%
YouGov	Wald	49.40%	52.60%	47.40%	50.60%
	Agressti-Coul	49.41%	52.60%	47.40%	50.59%
	Clopper-Pearson	49.40%	52.62%	47.38%	50.60%

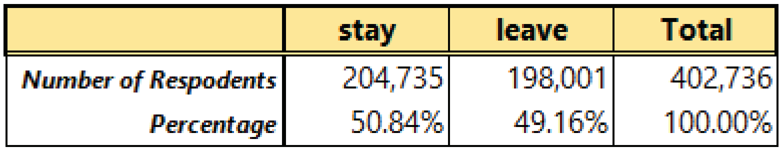
Figure3.4.1: *Box Plot-Range of Estimate (Clopper-Pearson CI)*



3.4.2. Combined Polls

Here, we include all historical polls and treat them as a single sample, weighting only by sample size. Basically, we assume there is no change in opinion of surveyed respondent over time. The point estimate and interval for the decided respondents (i.e. excluding the *Undecided*) are shown by Table3.4.4 and 3.4.5. Table3.4.4 exhibits the overall *Stay* is greater than the overall *Leave* variable, 50.84% over 49.16%. This outcome also is confirmed by Table3.4.5. There is a gap between interval of *Stay* and *Leave* with all 3 methods of CI provides equal range.

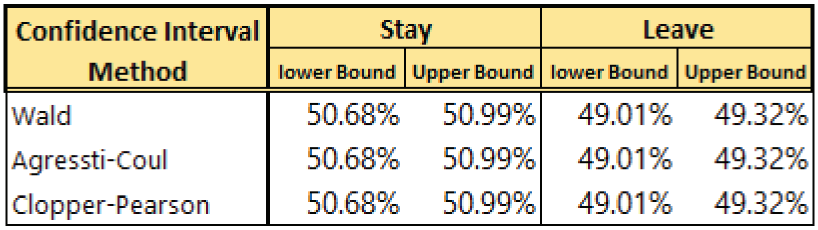
Table3.4.4: *Combined Polls Result – Point Estimate*



	stay	leave	Total
Number of Respodents	204,735	198,001	402,736
Percentage	50.84%	49.16%	100.00%

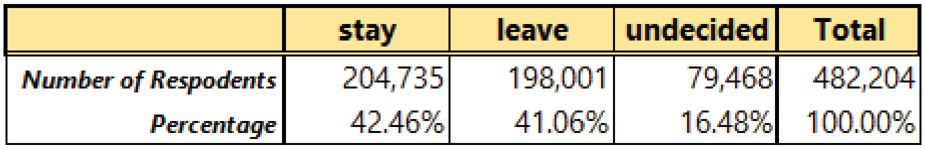
It is tempting to quickly declare that *Stay* will be the winner. However, these numbers exclude the huge percentage of respondents who did not decide their vote yet (the *Undecided*) as shown by Table3.4.6. On the actual voting day, most undecided people will vote either stay or leave. Thus, we need to pay attention on undecided respondents. In addition, how accurate is the assumption of no change in opinion of surveyed respondent over time? Is it true that time has no effect of public opinion EU referendum? Therefore, we will start analyzing the **trend of public opinion overtime** in next section.

Table3.4.5: *Combined Polls Result – Confidence Interval*



Confidence Interval Method	Stay		Leave	
	lower Bound	Upper Bound	lower Bound	Upper Bound
Wald	50.68%	50.99%	49.01%	49.32%
Agressti-Coul	50.68%	50.99%	49.01%	49.32%
Clopper-Pearson	50.68%	50.99%	49.01%	49.32%

Table3.4.6: *The Size of the Undecided Voters*



	stay	leave	undecided	Total
Number of Respodents	204,735	198,001	79,468	482,204
Percentage	42.46%	41.06%	16.48%	100.00%

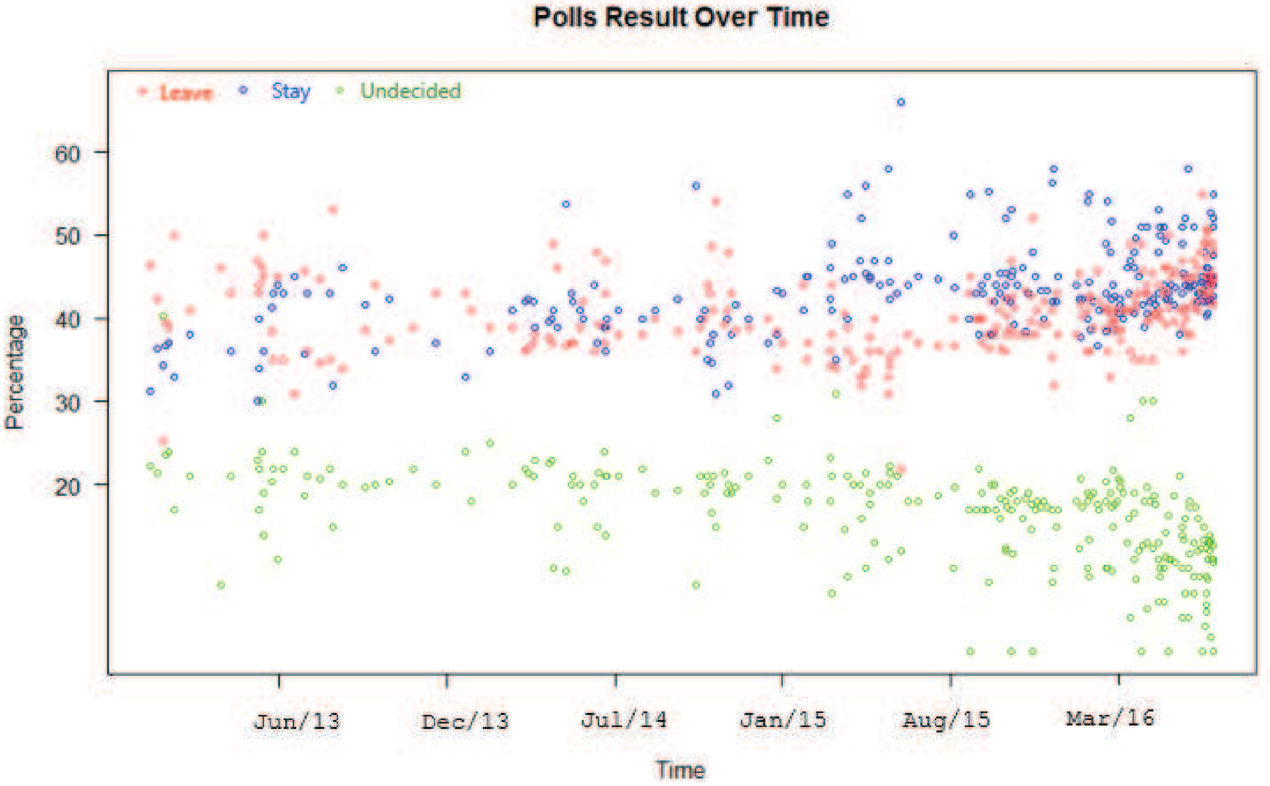
3.4.3. Weighted Polls

Here, although we include all historical polls, we will weighted them with help of kernel function. We consider nonparametric as the best approach since as shown in Figure3.4.2, trend of public opinion is clearly changes over time and not in linear relationship (This also suggest that *Combined Polls* may not the best technique to deal with pre-referendum polls data).

Based on visual observations from Figure3.4.2, start at **beginning of 2013** UK citizen seems favor "Leave" over "Stay". The trend slightly changes in 2014. During **mid-2015**, the public opinion showed contrast changed. We can obviously see that UK citizen favor "Stay" over "Leave". However, in **2016**, the trend changes again. Public opinion on "Leave" and "Stay" seems equal in the end of x-axis. In addition, we see that, as the data is approaching 23rd June 2016, the value of undecided is approaching zero. This is logical because respondents will more likely know their decision as time get closer to voting time.

Let us confirm this visual guess with a statistical inference which is appropriate for estimating proportional figures overtime with local nonparametric regression. Initially, we begin with Nadaraya-Watson Kernel Estimator.

Figure3.4.2: *Scatter Plots of Respondents Opinion Overtime*



We denoted s, l, u as variables *Stay*, *Leave*, *Undecided*. In this case, we want to estimate s, l, and u on time (t) using Nadaraya-Watson. Let $h > 0$ be a positive number called the bandwidth. The **Nadaraya-Watson (NW) kernel estimator** is defined by:

$$\hat{s}_n(t) = \sum_{i=1}^{n} \Pi_i(t) s_i \qquad \hat{l}_n(t) = \sum_{i=1}^{n} \Pi_i(t) l_i \qquad \hat{u}_n(t) = \sum_{i=1}^{n} \Pi_i(t) u_i \tag{3.1}$$

Where K is a kernel and Π as weight function for our case,

$$\Pi_i(t) = \frac{K\left(\frac{t - t_i}{h}\right)}{\sum_{j=1}^{n} K\left(\frac{t - t_j}{h}\right)} \tag{3.2}$$

We have aim that

$$\hat{s}_n(t) + \hat{l}_n(t) + \hat{u}_n(t) = \sum_{i=1}^{n} \Pi_\mathrm{h}(t) \{s_n(t) + l_n(t) + u_n(t)\} = 1 \tag{3.3}$$

Now, we can fit our data using Nadaraya-Watson kernel regression estimator as described in equation (3.3). One issue with Nadaraya-Watson is to determine the bandwidth h. Too low will yield overfitting estimation, too high will yield underfitting estimation. For instance, Figure3.4.3 and 3.4.4 show the fitting with Nadaraya-Watson under different bandwidth size.

Figure3.4.3 points that too high h give underfitting estimation. The fit is too smooth (almost look linear). On other hand, Figure3.4.4 shows that too low h give overfitting estimation. The fit is rough. Therefore, we need to find the "best" fitting with choosing the optimal bandwidth.

To solve this issue, we performed **cross-validation** to find the bandwidth which will minimize the errors. Cross-validation equations are defined below:

$$CV_s = \frac{1}{n}\sum_{i=1}^{n}\left(s_i - \hat{s}_{(-i)}(t_i)\right)^2 \qquad CV_l = \frac{1}{n}\sum_{i=1}^{n}\left(l_i - \hat{l}_{(-i)}(t_i)\right)^2 \qquad CV_u = \frac{1}{n}\sum_{i=1}^{n}\left(u_i - \hat{u}_{(-i)}(t_i)\right)^2 \tag{3.4}$$

Where $\hat{s}_{(-i)}, \hat{l}_{(-i)}$, and $\hat{u}_{(-i)}$ are the estimator obtained by omitting the i^{th} pair (t_i, s_i), (t_i, l_i), and (t_i, u_i) respectively.

Figure3.4.3: *Fitting Nadaraya-Watson Kernel Estimator with* $h = 180$

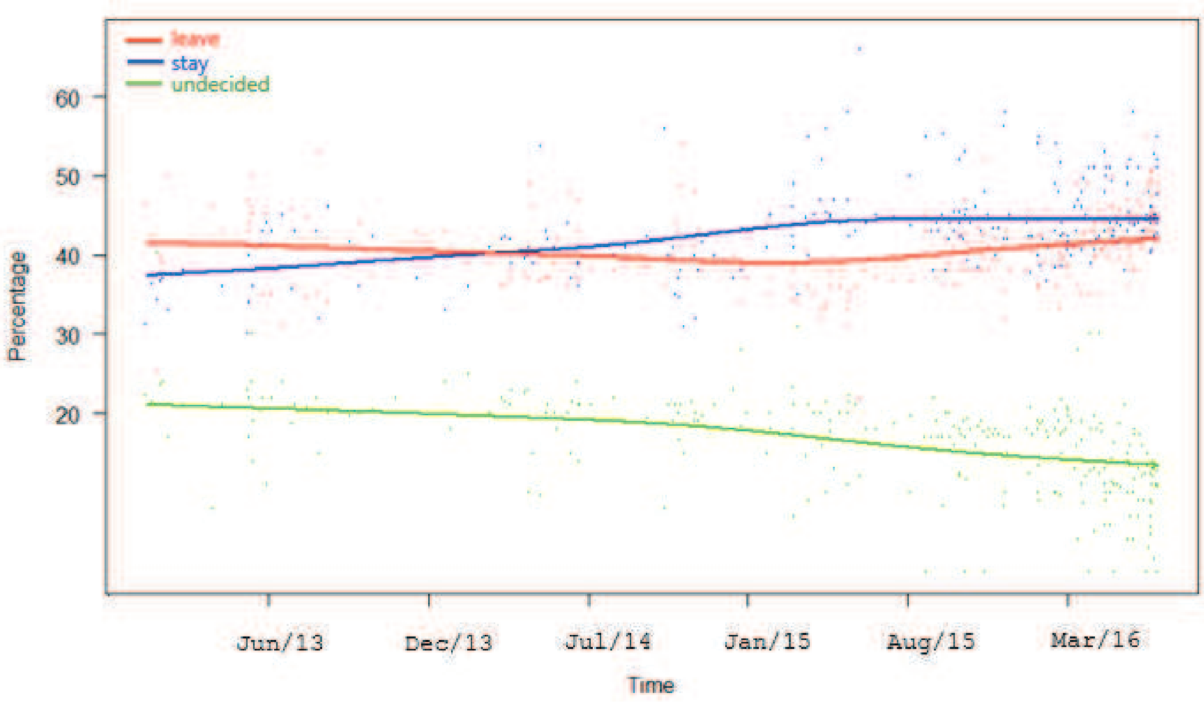
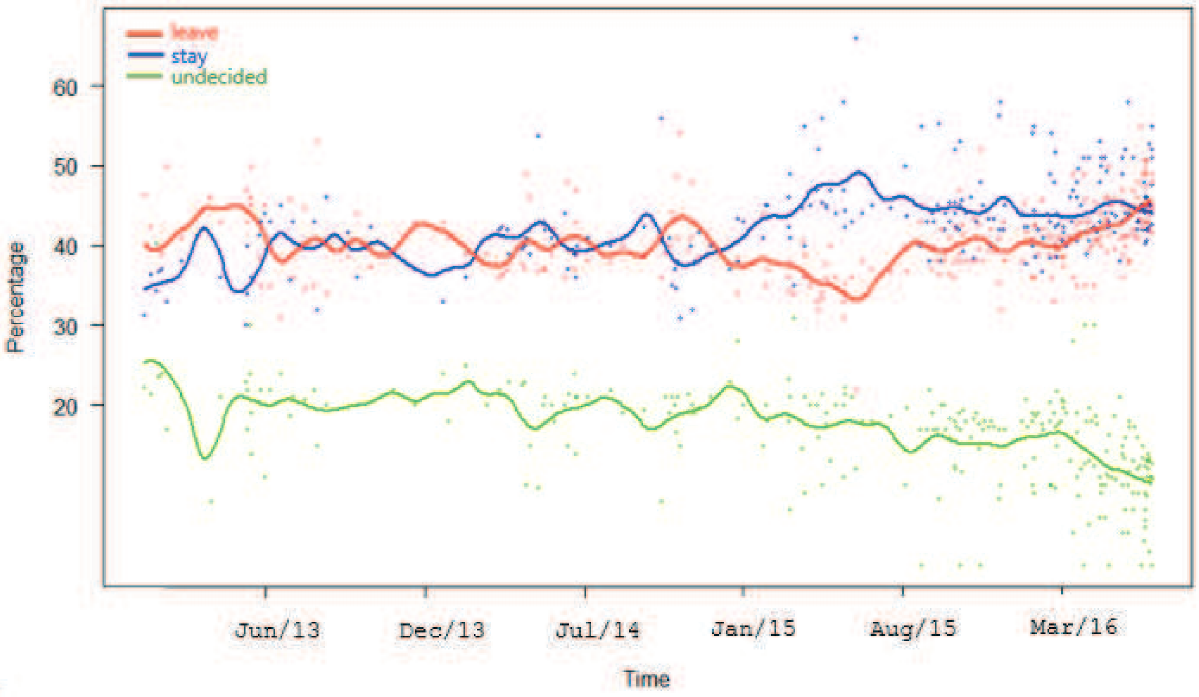
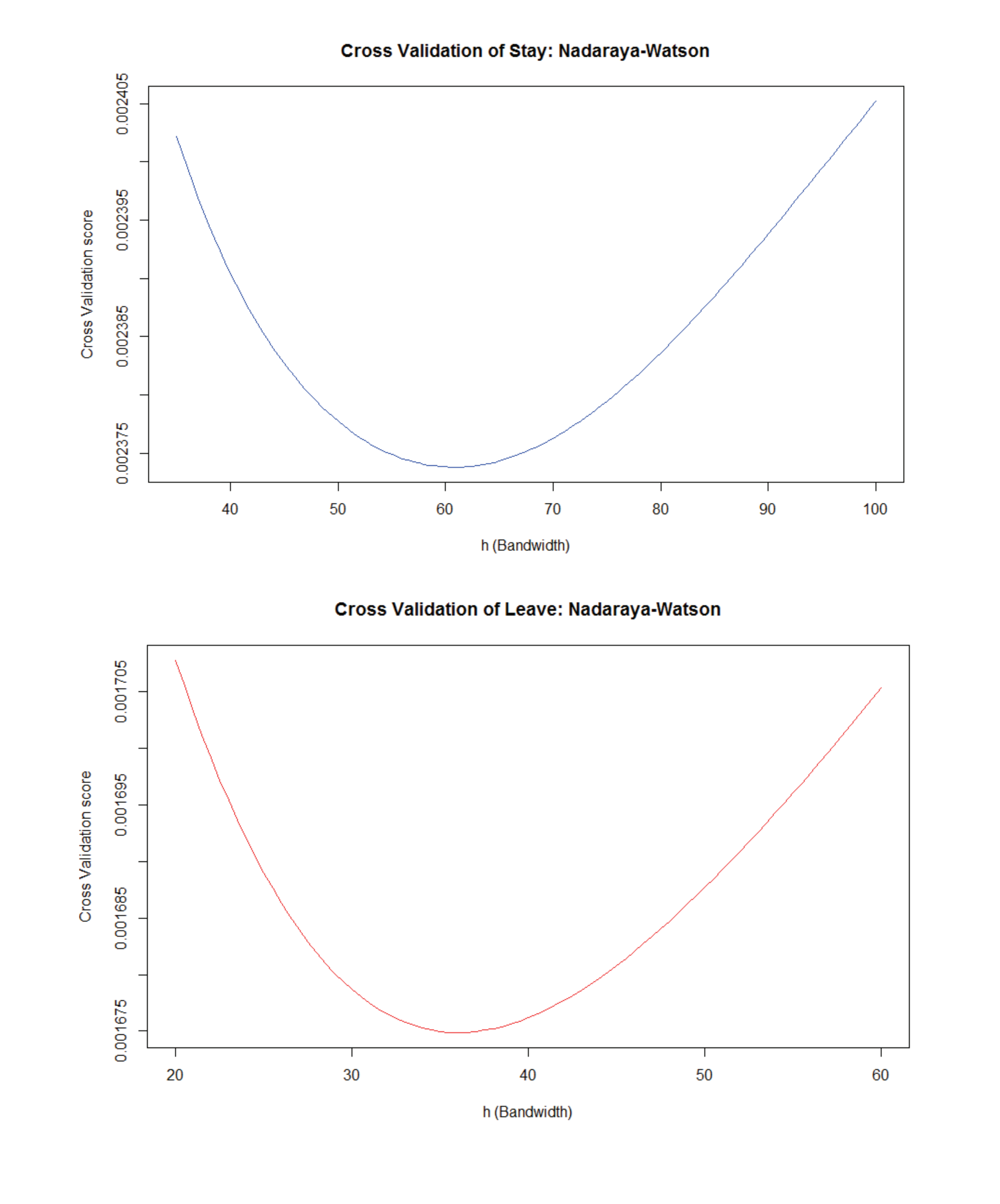


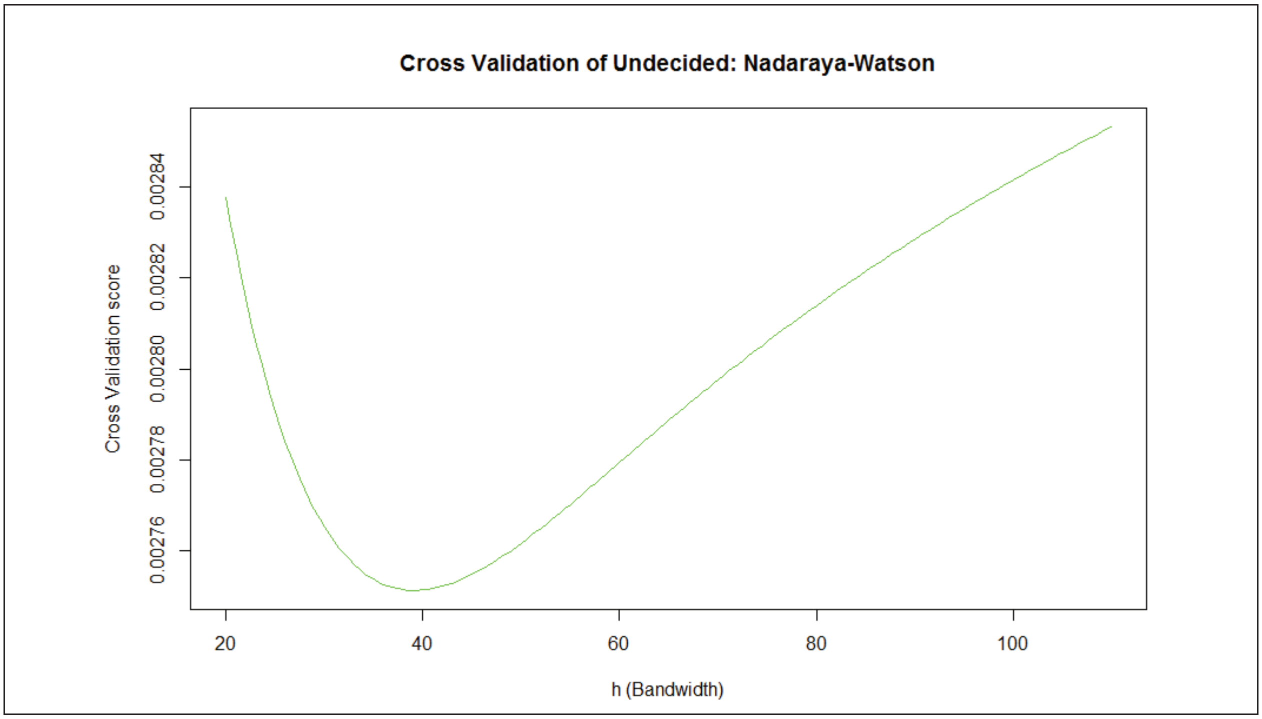
Figure3.4.4: *Fitting Nadaraya-Watson Kernel Estimator with* $h = 10$



Using cross-validation as described in equation (3.4) , we found that the most optimal bandwidth for *Stay*, *Leave*, and *Undecided* are: 60.93, 36.03, and 39.15 respectively. Figures 3.4.5 shows the plot of the cross-validation score versus bandwidth for *Stay, Leave* and *Undecided.*

Figure3.4.5: *Nadaraya-Watson Cross-Validation*



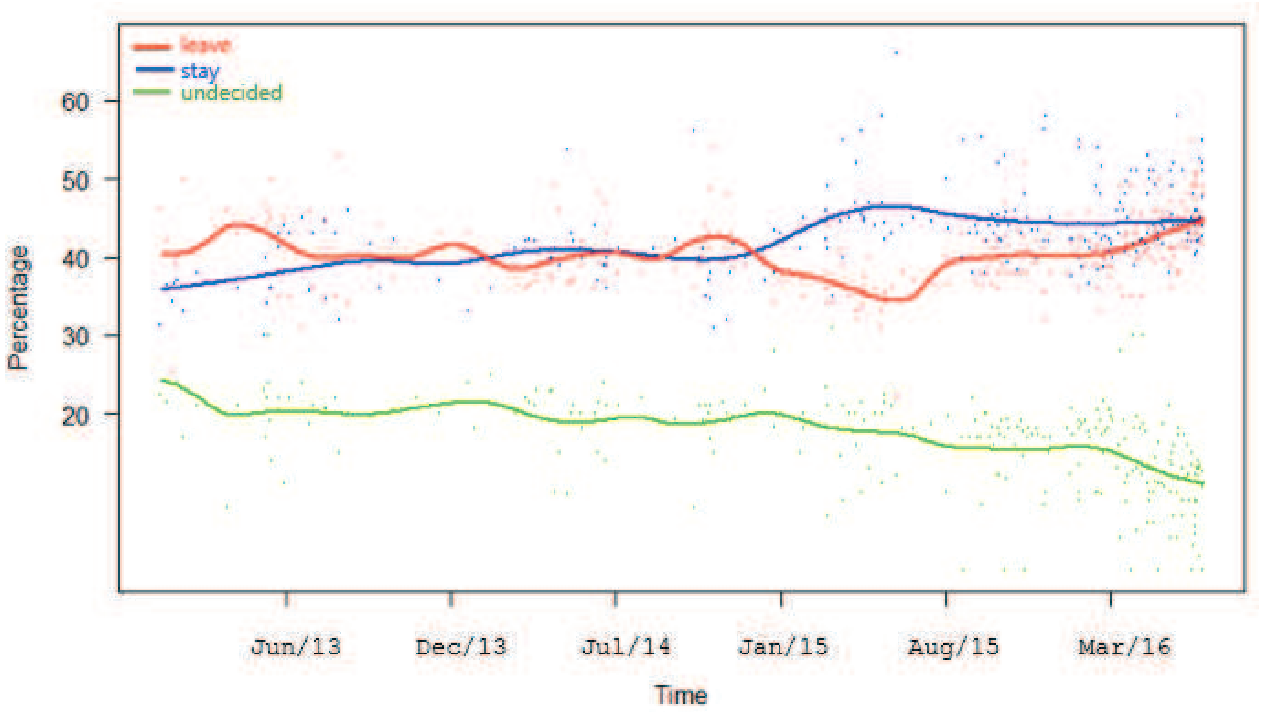


Now, we can use the most optimal bandwidths into our kernel regression estimator, we can compute the Nadaraya-Watson best fitting with the optimal bandwidth that we got previously. The result is shown by Figure3.4.6. From the graph, we can see the following patterns:

- Overtime, the number of *Undecided* is steadily declining.
- In early 2013, majority of respondents tended to vote *Leave*.
- Between late 2013 and 2014, there are swings of majority between *Leave* and *Stay*.
- Starting from early 2015, the *Stay* became leading majority. However, the *Leave* was steadily catching up until finally it surpassed *Stay* as the majority on 22nd June 2016.

These results confirm our initial guess based on observational. Now that we have the trends, can we use these trends for prediction? In other words, using the pre-referendum polling data to predict the actual referendum result. Next section will discuss the prediction.

Figure3.4.6: *Best fitting of Nadarawaya-Watson with minimum error bandwidth*



4. Predicting the UK EU Referendum Result

This section explores whether the Brexit referendum result is predictable from the pre-referendum polling data using nonparametric regression.

4.1. Predicting the Result using Local Polynomial Regression

To extend the trends, we will use **local polynomial regression** as it is more suitable than Nadaraya-Watson estimation. Nadaraya-Watson, can be considered as a simplified case of Local Polynomial Regression, suffers from boundary bias and design bias. In our case, boundary bias is a bias near the endpoints of the time (t_i), while design bias is a bias that depends on the distribution of the time (t_i). These problems can be alleviated by using a generalization of kernel regression called Local Polynomial Regression.

Now we want to find estimation of

$$\hat{s}_n(t) = \sum_{i=1}^{n} \varphi_i(t) s_i \qquad \hat{l}_n(t) = \sum_{i=1}^{n} \varphi_i(t) l_i \qquad \hat{u}_n(t) = \sum_{i=1}^{n} \varphi_i(t) u_i \tag{4.1}$$

Where

$\varphi_i(t)^T = (\varphi_1(t), \dots, \varphi_n(t)),$

$$\varphi_i(t)^T = e_1^T (T_t^T W_t T_t)^{-1} \, T_t^T W_t \tag{4.2}$$

$e_1 = (1,0, \dots ,0)^T$ and T_t and W_t are defined below:

$$T_t = \begin{pmatrix} 1 & t_1 - t & \cdots & \frac{(t_1 - t)^p}{p!} \\ 1 & t_2 - t & \cdots & \frac{(t_2 - t)^p}{p!} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & t_n - t & \cdots & \frac{(t_n - t)^p}{p!} \end{pmatrix}$$

and

W_t be $n \times n$ diagonal matrix whose (i, i) component is $w_i(t)$

Degree of polynomial denoted as p. If $p = 0$, then the estimation back again to Nadaraya-Watson. This local polynomial regression estimator has mean

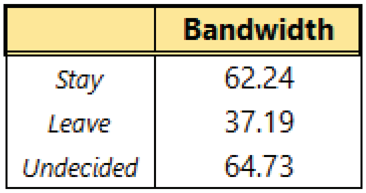
$$E(\hat{s}_n(t)) = \sum_{i=1}^{n} \varphi_i(t) s(t_i) \qquad E\left(\hat{l}_n(t)\right) = \sum_{i=1}^{n} \varphi_i(t) l(t_i) \qquad E(\hat{u}_n(t)) = \sum_{i=1}^{n} \varphi_i(t) u(t_i) \tag{4.3}$$

Here, we also aim to obtain that

$$\hat{s}_n(t) + \hat{l}_n(t) + \hat{u}_n(t) = \sum_{i=1}^{n} \varphi_{\mathrm{h}}(t)\,\{s_n(t) + l_n(t) + u_n(t)\} = 1 \tag{4.4}$$

Table4.1.1 shows the most optimal bandwidth using Local Polynomial Regression **degree=1** based on Cross-validation. As we aim to achieve condition as mentioned in equation (4.4), then we need to choose **one value of bandwidth** for all variables *Stay*, *Leave*, and *Undecided*. We decide to use bandwidth equal to 62 for two reasons. First, it close to each variables most optimal bandwidth and plot estimation looks fine (not rough and also not too smooth) as provided in Figure4.1.1.

Table4.1.1: *Most Optimal Bandwidth using Local Polynomial (degree=1) for Stay, Leave, and Undecided*



	Bandwidth
Stay	62.24
Leave	37.19
Undecided	64.73

Figure4.1.1: *Fitting Local Polynomial (degree=1, bandwidth=62)*

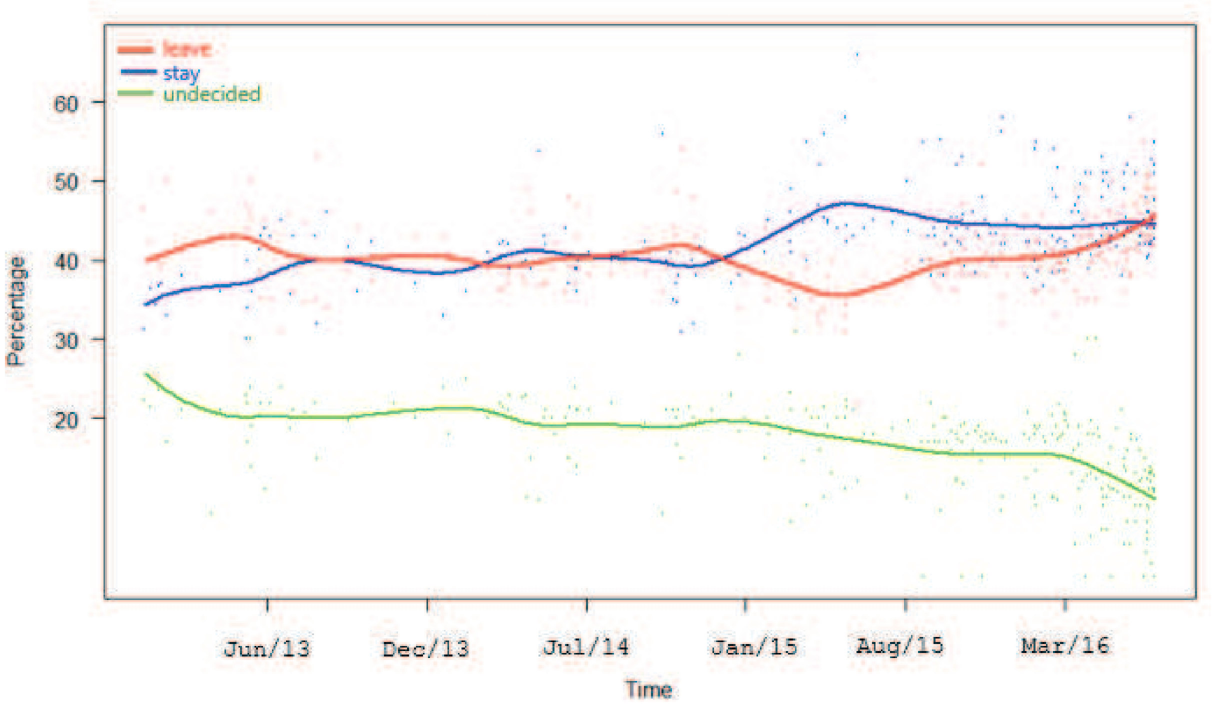


Figure4.1.1 shows the nonparametric regression fitting of local polynomial degree=1. We used help from "*np*", "*KernSmooth*", and "*npregfast*" package in R to compute this. Further R-code

details can be found in Appendix 3. Figure4.1.1 also confirms all our conclusion that we summarize in the end of section 3. In addition, at the end point of time t_i, Figure4.1.1 shows clearly that *Leave* is greater than *Stay*. However, to be precise, we will proceed to find confidence band for our fitting estimation.

Here, an approximate confidence band;

$$I(t) = \hat{r}_n(t) \pm c\hat{\sigma}(t)\left\|\varphi_i(t)\right\| \tag{4.5}$$

Where $\hat{r}_n$is the estimation variable. It can be either estimation for $s, l,$ or u. c is a solution of **tube formula.** More details on this topic can be found in Faraway and Sun (1995).

Figure4.1.2 provides the fitting estimation and confidence bands for three variables. We can see that even *Leave* is greater than *Stay* at end point of x-axis (22nd June 2016). However, their interval overlapping with each other. Now, the key question, how about on voting day? What is public opinion regarding Brexit in **23rd June 2016**? Therefore, we use the model to predict the public opinion. To increase accuracy, we will do a 95% confidence interval estimate in addition to point prediction.

Figure4.1.2: *Confidence Bands of Local Polynomial Regression degree=1, bandwidth=62*

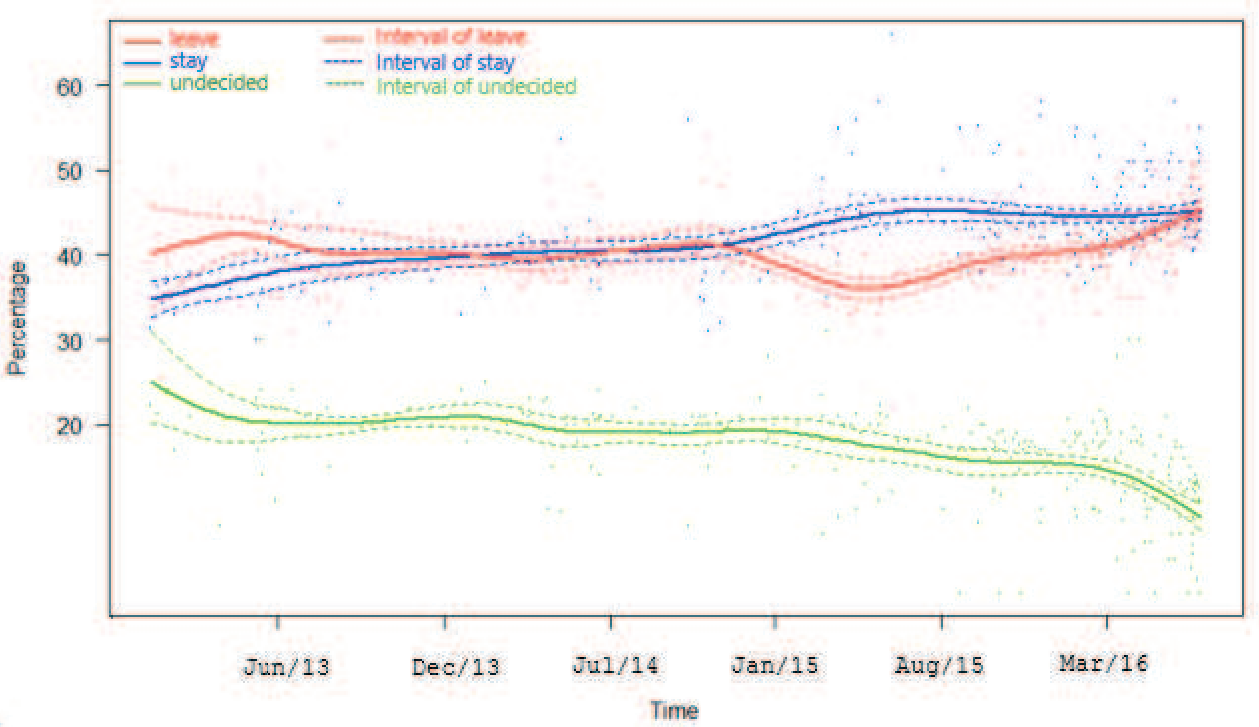
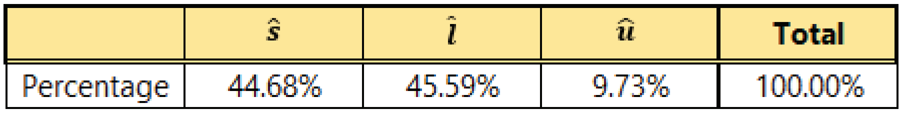


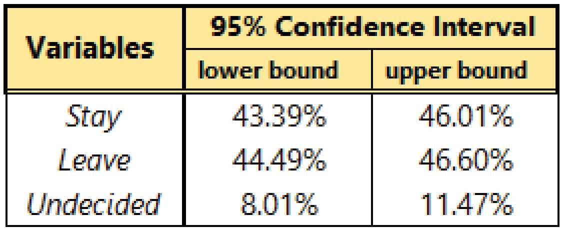
Table4.1.2 provides prediction results of public opinion in 23rd June 2016. We see that based on point estimate prediction that *Leave* is slightly greater than *Stay*. It means that with historical polls data, there is evidence if majority of UK citizen prefers to leave EU. Our prediction is also exactly equal to 100% as our client's demand. However, to be sure with our prediction, we should check the 95% confidence interval. Table4.1.3 points the outcome of 95% confidence interval. We clearly can see that there is overlapping interval between *Stay* and *Leave.* Thus, we can conclude based on analysis on historical polls on pre-referendum, Brexit is **not surprising** event.

Table4.1.2: *Point Estimate of Prediction for Brexit Result*



	$\hat{s}$	$\hat{l}$	$\hat{u}$	**Total**
Percentage	44.68%	45.59%	9.73%	100.00%

Table4.1.3: *95 % Confidence Interval of Prediction for Brexit Result*



Variables	**95% Confidence Interval**	
	lower bound	**upper bound**
Stay	43.39%	46.01%
Leave	44.49%	46.60%
Undecided	8.01%	11.47%

The insights from the prediction result are as follow:

- The model indeed predicts the Brexit Referendum would result in *Leave* as majority with very small difference. This finding suggests that Brexit is more likely to happen than not.
- The Prediction interval between *Leave* and *Stay* are overlapping. This finding clearly suggests that Brexit is not a shocking result. The prediction based on polls data has shown that it would be a very close call.

One thing we see from the result is that the number of *Undecided* is still quite high. Since the gap between *Leave* and *Stay* is quite narrow, the last minutes swing from the undecided may change the result of the referendum. The *Undecided* in the polls clearly creates big uncertainty in this case. For this reason, we will perform maximum uncertainty analysis. Next, we will compare our result with one of previous study and do maximum uncertainty analysis.

4.2. Comparison Result with Previous Studies

In this section, we will compare the results that we got with previous study from John Fry (2016). The study claimed claim that the Brexit referendum is not predictable as simple linear regression prediction would ultimately turn out to be wrong.

We will compare our local polynomial regression model with simple linear regression model. The result is given by Figure4.2.1. Linear regression is less sensitive to changes and the fitting estimation clearly shows that *Stay* is greater than *Leave* at end of x-axis. While the Local Polynomial regression is much more sensitive to latest changes and the fitting estimation result is in contrast at end of x-axis.

Table4.2.1 provide outcome of prediction on voting day with linear regression. The point-prediction indeed shows that *Stay* greater than *Leave.* Nevertheless, the interval has overlapping. Thus, previous study's claim of Brexit is not obvious from poll data may be incorrect. In addition, there are two drawbacks from using **simple linear regression** for our Brexit case. First, the data is clearly not in linear relationship. Forcing to use linear regression on this case will give us two main consequence: the point prediction is apparently less sensitive compare nonparametric regression and the prediction interval from linear regression is obviously too wide. It is hard to inference conclusion. Second, the prediction of three variables are not equal to 1. This maybe can be solved with re-weighting.

Figure4.2.1: *Comparison Fitting between Local Polynomial vs Linear Regression*

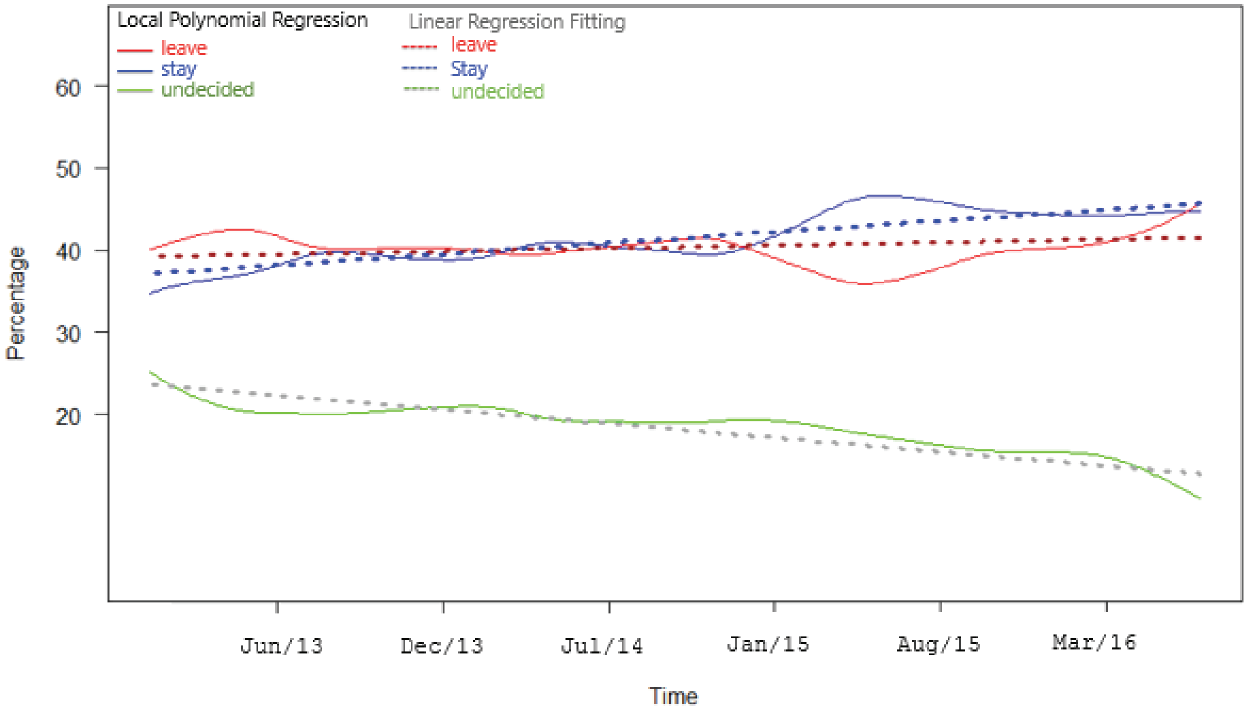
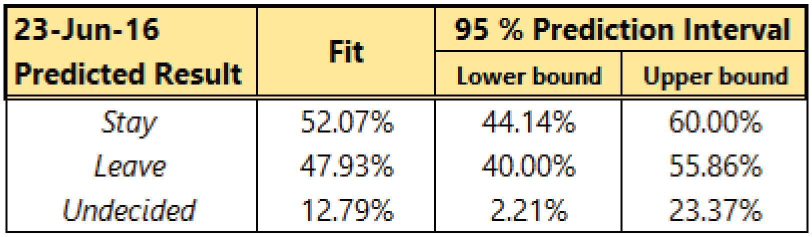


Table4.2.1: *Prediction Result with Linear Regression*



23-Jun-16 Predicted Result	Fit	95 % Prediction Interval	
		Lower bound	Upper bound
Stay	52.07%	44.14%	60.00%
Leave	47.93%	40.00%	55.86%
Undecided	12.79%	2.21%	23.37%

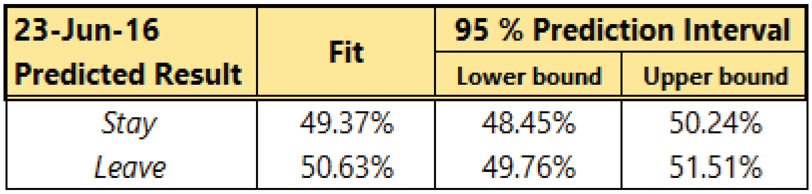
4.3. Maximum Uncertainty Analysis

To account for the uncertainty from the *Undecided*, we will need to allocate the proportion of *Undecided* into *Stay* and *Leave.* To take account of account maximum uncertainty, we will assume 50:50 split of *Undecided* for both *Stay* and *Leave.* For example, a poll with following result: 40% *Stay*, 50% *Leave*, and 10% *Undecided*; will become 45% *Stay* and 55% *Leave.* We split for all 261 observations then proceed to analysis.

Next, using Local Polynomial Regression degree=1 and bandwidth = 48.83 (based on Cross-validation, h_{stay} = 48.83 and h_{leave} = 48.83 are the optimal bandwidth), we got outcome as shown in Figure4.3.1 At end of x-axis of Figure4.3.1, we clearly see that *Leave* is greater than *Stay*. Then, we proceed to observe the confidence band and prediction interval for further confirmation.

Figure4.3.2 shows confidence band for fitting in maximum uncertainty analysis. At the end of June 2016, the interval between *Leave* and *Stay* still overlapping. Table4.3.1 provides prediction result for real voting day. From the table, we can see that the model predicts that *Leave* would win the referendum result. We also see that the CI ranges have become narrower. Furthermore, the CI overlap between *Leave* and *Stay* has reduced.

Table4.3.1: Prediction Result Based on Maximum Uncertainty Analysis



23-Jun-16 Predicted Result	Fit	95 % Prediction Interval	
		Lower bound	Upper bound
Stay	49.37%	48.45%	50.24%
Leave	50.63%	49.76%	51.51%

Figure4.3.1: *Local Polynomial Regression Model with Maximum Uncertainty Split*

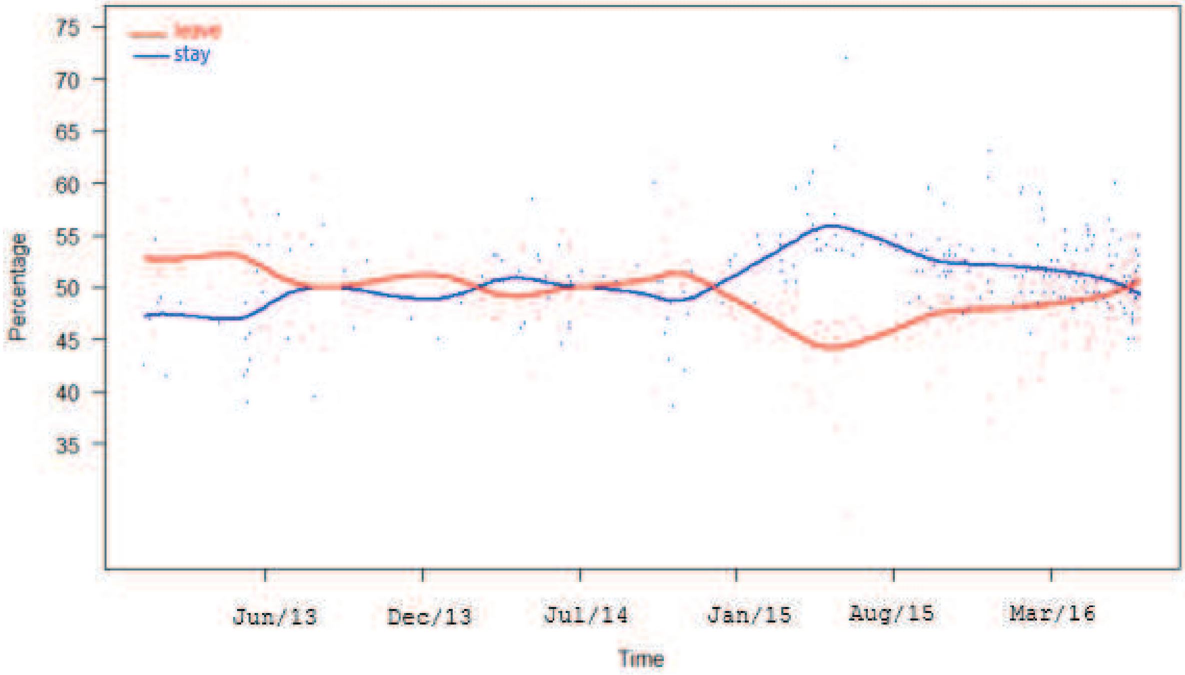
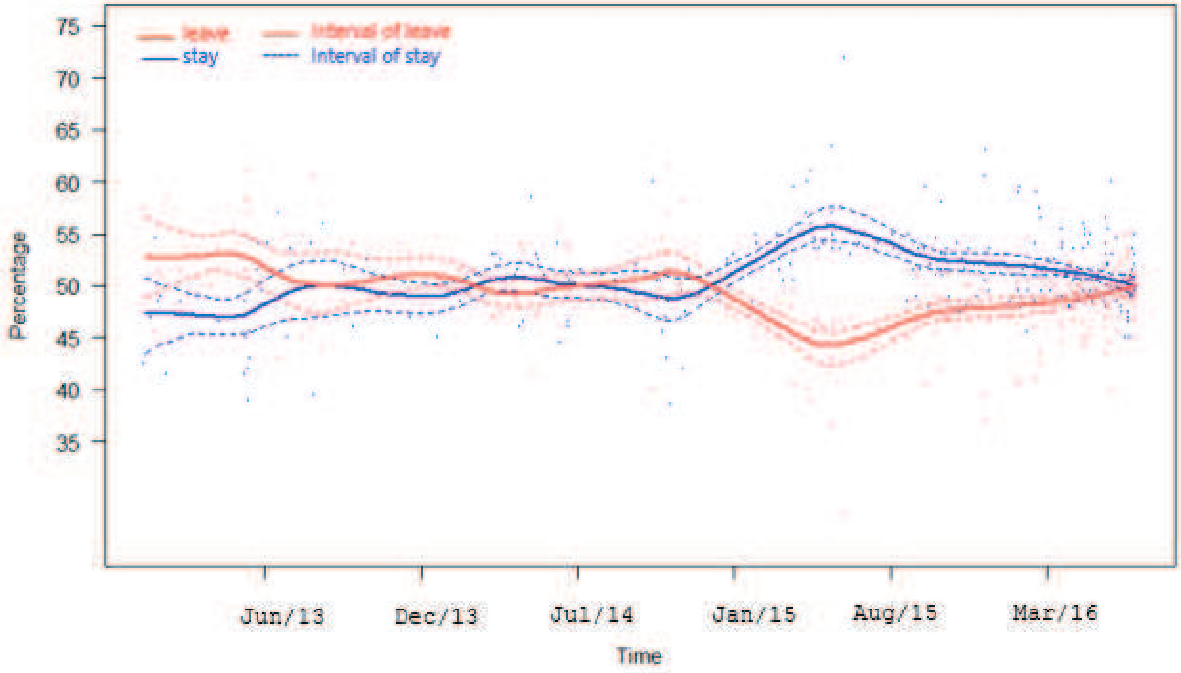


Figure4.3.2: *Local Polynomial Regression Model with Maximum Uncertainty Split* with Confidence Bands



5. Conclusion

This section summarises the findings of this study, provides recommendations as well as potential further studies not yet covered by this study.

5.1. Conclusion

Our analysis on **Section 3 and 4** suggests that:

1. Given the overlapping confidence intervals between *Leave* and *Stay,* the referendum result of majority *Leave* is actually not a shocking result. Many pollsters made prediction mistake because they only rely on point estimate.
2. The Brexit is actually not a surprising result. The prediction based on polls data has shown that it would be a very close call. Unfortunately, one needs to use more robust and sophisticated methodologies to extract the insights from the patterns. Many pollsters made prediction mistake because they rely on descriptive analysis and latest poll result only. They do not consider the trend over time, they do not model the historical trend using non-parametric regression. They ignore the *Undecided* which is a big source of uncertainty in this case.
3. Despite its non-trivial bias, the historical polls data still can be trusted. It is just that one needs to aware of this bias and quantify the uncertainty in the prediction.

5.2. Recommendation

Going forward, to ensure better predictive accuracy, we suggest the pollsters to:

- Always consider the confidence interval when making prediction. Do not rely only on the point estimate.
- Do not ignore the *Undecided* when their proportion is big and can swing the result easily.
- Analyse trends overtime, do not rely on latest polls only. They may uncover some insights.
- Use non-parametric estimation/regression when it is more appropriate to do so. Linear regression, despite its simplicity, is not the best tool for all cases.

5.3. Caveats and Further Improvement

This study made the following assumptions which may not be valid.

1. The pollsters used relatively similar sampling methodology. In reality, we are not sure about this. A further study on different methodologies used and their accuracy will be useful.
2. The pollsters followed robust sampling technique to ensure sampling randomness and representativeness. In reality, this may not be true. For example, there may be demographic

bias in the way voters are polled. Some pollsters were using online polling. Older voters who tended to vote for "Leave" are less likely to reply to online polling. Some pollsters were using telephone polling which includes high percentage of young people. Although the younger voters tended to support *Stay*, their turn out was lower, thus online polling skewing the result towards *Stay*. Therefore, a further study reviewing the robustness of the sampling methodology is needed.

3. The polling results between different pollsters are independent and identically distributed. This may not be true. Perhaps, there is a herd mentality and the result of one pollsters affected the result of other pollsters. A further study investigating this assumption will be interesting.
4. The polling results are not auto-correlated. But in reality the result of previous polls may affecting the result of later polls. A further study on this will be reveal more insights.

Words count: 4,971

References

Bergman, J & Holmquist B. 2014. “Poll of Polls: A Compositional Loess Model”. Scandinavian Journal of Statistics, page 301-302 [BH14]

Brown, L.B. & Chappell H.W.Jr. 1999. “Forecasting Presidential Election using History and Polls”. International Journal of Forecasting No 15, page 127-135. [BC99]

Campbell, J. E. 1996. “Polls and Votes: The Trial Heat Presidential Election Forecasting Model, Certainty, and Political Campaigns”. American Politics Quarterly No 24, page 408-433. [C96]

Clarke, H & Goodwin, M. 2016. “Why Britain Voted for Brexit: An Individual-Level Analysis of the 2016 Referendum Vote” [CG16]

Campbell, J.E. & Wink K. A. 1990. “Trial-Heat Forecasts of the Presidential Vote”. American Politics Quarterly No 18, page 251-269. [CW90]

Celli, Fabio et al. 2016. "Predicting Brexit: Classifying Agreement is Better than Sentiment and Pollsters". Proceedings of the Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media, pages 110-118. Osaka, Japan, December 2016. [CF16]

Christensen, W. F. & L.W. Florence. 2008. “Predicting Presidential and Other Multistage Election Outcomes Using State-Level Pre-Election Poll”. The American Statistician, 62:1, page 1-10. DOI: 10.1198/000313008X267820. [CW08]

Fisher, S. D. 2016. "Putting it all together and forecasting who governs", Electoral Studies No 41, page 234-238. [FS16]

Fisher, S. D. and Lewis-Beck, M. S. 2016. "Forecasting the 2015 British General Election". Electoral Studies No 41, page 225-229. [FLB16]

Fry, J. M. 2014. "Statistical Prediction of the Scottish Referendum". August 2014. [F14]

Fry, John. 2016. "A Statistical Reaction to Brexit". August 2016. [F16]

Johnston, Ron et al (A). 2016. "Can We Really Not Predict Who Will Vote for Brexit, and Where?”.March 2017. [JRA16]

Johnston, Ron et al (B). 2016. "Predicting the Brexit Vote: Getting the Geography Right (More and Less)". February 2017. [JRB16]

Simon, Jenkins. 2016. On Brexit, gender, age and political party. Accessed on 15th August 2017. [JS16]

Sturgis, P. et al. 2016. "Report of the Inquiry into the 2015 British General Election Opinion Poll". Accessed on 15th August 2017. [P16]

Appendix

Appendix 1 - Projection of a point on a plane

Let us consider plane p with equation, $ax + by + cz + d = 0$ and a point $M(u, v, w)$.

Thus, we can calculate normalization line by

$$\frac{x-u}{a} = \frac{y-v}{b} = \frac{z-w}{c} = t$$

We continue to get a point on the line parametric equations by

$$x = u + at \qquad y = v + bt \qquad z = w + ct$$

Then, we can compute the value of t_0 with a point of this line on the plane

$$a(u + at_0) \ + b(v + bt_0) + c(w + ct_0) + d = 0$$

$$t_0 = -\frac{au + bv + cw + d}{a^2 + b^2 + c^2}$$

Let us denote x_0, y_0, z_0 as projection of point $M(u, v, w)$

So we have the projection of u, v, w:

$$x_0 = u + at_0 \qquad y_0 = v + bt_0 \qquad z_0 = w + ct_0$$

Appendix 2 – Confidence Interval of Binomial Proportion

A popular confidence interval for a binomial proportion, p, is the **Wald CI**. The equation for 95% confidence is:

$$\hat{p} \ \pm 1.96 \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Although such approximate confidence intervals are implemented in many statistical packages—and widely used in practice—their performance has been heavily criticized. There is however an interval with the same form, but with a different center $\tilde{p}$, and a modified value for n, which is known to have better coverage. The modified interval are known as **Agresti-Coull**:

$$\tilde{p} \pm 1.96 \sqrt{\frac{\tilde{p}(1-\tilde{p})}{\tilde{n}}}$$

Where

$$\tilde{p} = \ \tilde{X}/\tilde{n} \qquad \tilde{n} = n + 1.96^2 \qquad \tilde{X} = X + \frac{1.96^2}{2}$$

Another popular method is the **Clopper–Pearson** interval:

$$CI_{\alpha}^{CP} = (B(a/2\,; X, n - X + 1), B(1 - a/2\,, X + 1, n - X))$$

While B is the quantile of beta distribution.

Appendix 3 – R Code

```
library(np)

library(PropCIs)

library(MASS)

library(KernSmooth)

dta <- read.csv('data2.csv')

t0 <- -(dta$stay+dta$leave+dta$undecided-1)/3

stay.p        <-dta$stay       +t0

leave.p       <-dta$leave      +t0

undecided.p   <-dta$undecided  +t0

total.p       <-stay.p + leave.p + undecided.p


plot(dta$time,stay.p,main="Polls Result Over Time",xlab="Time", ylab="Percentage",ylim=c(0,
0.67), yaxt="n",col="blue",cex=.3)

axis(2, at=pretty(stay.p), lab=pretty(stay.p) * 100, las=TRUE);
points(dta$time,leave.p,col="firebrick1",cex=.3); points(dta$time,undecided.p,
col="chartreuse3",cex=.3)

lines(locpoly(dta$time,stay.p     , kernel = "normal", bandwidth = 60, degree =
1),col="blue",lty=1,lwd=2)

lines(locpoly(dta$time,leave.p    , kernel = "normal", bandwidth = 60, degree = 1),col="red"
,lty=1,lwd=2)

lines(locpoly(dta$time,undecided.p, kernel = "normal", bandwidth = 60, degree =
1),col="chartreuse3" ,lty=1,lwd=2)


# choose h by cross validation

CrossValid <- function(h, xi, yi, LP.order = 3){

  n = length(xi);  z =0;  m=0

   for(i in 1:n) {

    subs = setdiff(1:n, i)

    fitLP = locpoly(xi[subs], yi[subs], kernel = "normal", bandwidth = h, degree = LP.order)

    ind = findInterval(xi[i], fitLP$x)

    if(length(ind) >= 1) {z = z + sum((yi[i] - fitLP$y[ind])^2); m = m+1}

  }

  return(z/m)

}

hs.s  = seq(35,100,0.5) ; hs.l  = seq(20,60,0.5);hs.u  = seq(20,110,0.5)

CVscore.s  = rep(NA, length(hs.s));CVscore.l  = rep(NA, length(hs.l));CVscore.u  = rep(NA,
length(hs.u))

for(i in 1:length(hs.s)){

  CVscore.s[i] = CrossValid(hs.s[i], dta$time,stay.p, LP.order = 1)

  CVscore.l[i] = CrossValid(hs.l[i], dta$time,leave.p,LP.order = 1)

  CVscore.u[i] = CrossValid(hs.u[i], dta$time, undecided.p, LP.order = 1)
}
#  h that minimises CV?
hcv.s = optimise(CrossValid,interval=c(5,130), xi=dta$time,
                 yi=stay.p, LP.order=1)$minimum
hcv.l = optimise(CrossValid,interval=c(5,130), xi=dta$time,
                 yi=leave.p, LP.order=1)$minimum
hcv.u = optimise(CrossValid,interval=c(25,130), xi=dta$time,
                 yi=undecided.p, LP.order=1)$minimum
np1 <- npreg(s  ~ time,regtype = "ll",bws= 45,p=1,bwmethod = "cv.aic",gradients= TRUE,data = df)
np2 <- npreg(l  ~ time,regtype = "ll",bws= 45,p=1,bwmethod = "cv.aic",gradients= TRUE,data = df)
np3 <- npreg(u  ~ time,regtype = "ll",bws= 45,p=1,bwmethod = "cv.aic",gradients= TRUE,data = df)
## With npregfast library
#-------------------------------------------------
library(npregfast)
fit1 <- frfast(s ~ time, data = df, nboot = 500, model="np",smooth="kernel",
kernel="gaussian",p=1, h0=0.05)
fit2 <- frfast(l ~ time, data = df, nboot = 500, model="np",smooth="kernel",
kernel="gaussian",p=1, h0=0.05)
fit3 <- frfast(u ~ time, data = df, nboot = 500, model="np",smooth="kernel",
kernel="gaussian",p=1, h0=0.05)
stay.es      <- predict(np1, newdata = cps.eval,interval="prediction")
leave.es     <- predict(np2, newdata = cps.eval,interval="prediction")
undecided.es <- predict(np3, newdata = cps.eval,interval="prediction")
stay2.est      <- predict(fit1, newdata = cps.eval,interval="prediction")
leave2.est     <- predict(fit2, newdata = cps.eval,interval="prediction")
undecided2.est <- predict(fit3, newdata = cps.eval,interval="prediction")
```