

UK voter preferences on the EU referendum

Mickey Warner
AMS 207 Quiz 2

Abstract

On 23 June 2016, the United Kingdom will hold a referendum to decide whether the nation should leave or remain in the European Union. Beginning September 2015, a variety of online and phone polls have been conducted to measure public preference on the issue. We consider regression models with linear time trends and additive effects for the polling source and for the type of poll (online or phone). We find evidence supporting a time trend with small effects due to the sources and the type of poll.

1. Introduction

Leading up to the 23 June 2016 referendum, over the course of nine months, starting September 2015, nine sources have published a total of one hundred polling results. The polls inquired voter preference on whether the United Kingdom (UK) should leave or remain in the European Union (EU). The possible choices for the respondents were “leave”, “remain”, and “don’t know.” We will be interested in the proportion of people who prefer to leave given that an opinion was given (i.e., we ignore the “don’t know” group).

Respondents giving “don’t know” as their preference accounted for about 16% of the voters in each poll (on average). Such a response may provide some information for preference on the referendum, but when it comes to voting time, a vote of “I don’t know” will not contribute to a decision.

The surveys came in two forms, either online or by phone. The majority of polls were online (78%) while the rest were phone (22%). Most sources were favored one kind of poll over the other. Only one source (Survation) had roughly half online (6 of 10) and half phone (4 of 10). Two other sources (ComRes and ICM) provided surveys of both types, but ComRes only 1 of 10 polls online and ICM had 35 of 37 online. The remainder of the sources provided only one type of poll.

A few of the sources published very few polls: BMG did 5, Opinium did 3, and Panelbase did 1 (all online). These three sources were combined into the Other category. By combining them, we are assuming they each provide the same effect on the outcome of the polls (i.e. their methodology is biased in the same way). This assumption may not be valid, but it allows us to obtain an estimate of the source effect which may not be possible with such few observations. The six sources that were not combined contributed from 7 to 37 of the polls.

Our focus in this paper will be to assess three aspects of the data: (1) whether there is a linear time trend in pref-

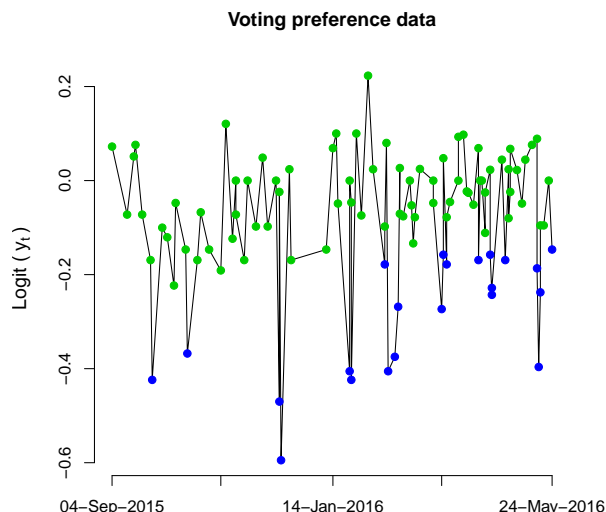


Figure 1: The logit of voting preferences over time. Green dots are online polls and blue dots are phone polls.

erence, (2) whether there are differences due to the various sources, and (3) whether there is a difference due to the polling type. A time-series model would seem a natural choice with this data. However, a regression model, with time as a covariate, is perhaps an easier model to work with in analyzing our three questions of interest. As such, we will work with regression models in this paper.

In section 2, we outline the model and methods used in the analysis. Results are reported in section 3 and we conclude with a discussion in section 4.

2. Methods

We define our $n = 100$ response variables as the logit of the ratio of leave percentage over leave plus remain percentage:

$$\begin{aligned} y_i &= \text{logit} \left(\frac{l_i}{l_i + r_i} \right), \quad i = 1, \dots, n \\ &= \text{logit}(p_i) \\ &= \log \left(\frac{p_i}{1 - p_i} \right) \end{aligned}$$

This transformation puts our response variable on the real line, allowing us to model y_i with a normal distribution. Each observation occurred at time t_i , which represents the number of days from 1 Sep 2015 (e.g. the first

Table 1: Design matrices \mathbf{X} for the eight models

Model	\mathbf{X}
M1	$(\mathbf{1}_n)$
M2	$(\mathbf{1}_n, \mathbf{t})$
M3	$(\mathbf{1}_n, \mathbf{S})$
M4	$(\mathbf{1}_n, \mathbf{P})$
M5	$(\mathbf{1}_n, \mathbf{t}, \mathbf{S})$
M6	$(\mathbf{1}_n, \mathbf{t}, \mathbf{P})$
M7	$(\mathbf{1}_n, \mathbf{S}, \mathbf{P})$
M8	$(\mathbf{1}_n, \mathbf{t}, \mathbf{S}, \mathbf{P})$

observation is $t_1 = 4$ as it occurs on 4 Sep 2015). Figure 1 shows the plot of (t_i, y_i) for $i = 1, \dots, n$ but with the x-axis relabeled to show the date corresponding to the t_i value. A quick glance shows difference between the two types of polls.

2.1 Regression model

Though we will consider eight models in this paper, they each have the general form

$$\mathbf{y} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I})$$

where $\mathbf{y} = (y_1, \dots, y_n)^\top$. The difference is in the design matrix \mathbf{X} . As discussed in the introduction, we wish to explore three aspects of the data: (1) whether there is a linear trend, (2) an effect due to source, and (3) an effect due to poll type. We will build design matrices for each combination, making in total $2^3 = 8$ models to consider.

Let $\mathbf{t} = (t_1, \dots, t_n)^\top$ be the vector of days since 1 Sep 2015.

Let \mathbf{S} be the $n \times 6$ binary matrix indicating the source to which the observations belong. That is, for row i , the j element is 1 if observation i belongs to source j and 0 otherwise. There are six columns in \mathbf{S} since after combining three of the sources, we were left with seven groups. For the design matrix to be full-rank, we must drop one of the columns of \mathbf{S} . Here, we choose to drop the Other category, so it will act as a baseline.

Let \mathbf{P} be the length n binary vector, which element i equal to 1 is observation i was a phone survey, and 0 otherwise. This means the online type is our baseline. The eight models each include an intercept term and are constructed by appending combinations of \mathbf{t} , \mathbf{S} , and \mathbf{P} . The design matrix \mathbf{X} for each model is given in Table 1. By fitting and comparing each model, we can get an idea of whether the components are important in modeling voting preference. The metrics used to compare the models are given in 2.4. Additionally, we may inspect the posterior distribution for $\boldsymbol{\beta}$ and see which components contain zero in a credible interval.

2.2 Priors for $\boldsymbol{\beta}$ and σ^2

We use a Zellner’s g -prior for $\boldsymbol{\beta}$ and a non-informative prior for σ^2 :

$$\begin{aligned} \boldsymbol{\beta} | \sigma^2 &\sim N(\mathbf{m}, g\sigma^2(\mathbf{X}^\top \mathbf{X})^{-1}) \\ p(\sigma^2) &\propto (\sigma^2)^{-1} \end{aligned}$$

where \mathbf{m} is chosen as the zero vector and g is fixed at n . This specification produces the following posterior full conditionals

$$\begin{aligned} \boldsymbol{\beta} | \sigma^2, \mathbf{y}, \mathbf{X} &\sim N\left(\frac{1}{g+1}(g\hat{\boldsymbol{\beta}} - \mathbf{m}), \frac{g\sigma^2}{g+1}(\mathbf{X}^\top \mathbf{X})^{-1}\right) \\ \sigma^2 | \boldsymbol{\beta}, \mathbf{y}, \mathbf{X} &\sim IG\left(\frac{n}{2} + \frac{p}{n}, \frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \frac{1}{2g}(\boldsymbol{\beta} - \mathbf{m})^\top (\mathbf{X}^\top \mathbf{X})(\boldsymbol{\beta} - \mathbf{m})\right) \end{aligned}$$

where $\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$ is the maximum likelihood estimate for $\boldsymbol{\beta}$ and p is the number of columns of \mathbf{X} (which changes from model to model). Thus we have closed-form full conditionals and we can explore the posterior using Gibbs sampling.

2.3 Model validation

We utilize two measures for model fit. The first is a χ^2 test from Johnson (2004) in “A Bayesian χ^2 Test for Goodness-of-Fit.” The test is comparable to classical χ^2 goodness-of-fit tests. With this test, we calculate a distribution of p -values where high values are in favor of a good model fit. See Johnson (2004) for details.

The second measure is based on leave-one-out cross-validation. We obtain B posterior draws for each model while leaving one observation out, y_i . Posterior predictions $y_{j,b}^*$ are then obtained and we compute

$$p_{j,i} = \frac{1}{B} \sum_{b=1}^B \mathbb{1}(y_{j,b}^* \leq y_i), \quad j = 1, \dots, 8$$

Distribution theory requires that $p_{j,\cdot}$ be uniformly distributed for the model to be appropriate. The Kolmogorov-Smirnov (K-S) test is used to determine whether $p_{j,\cdot}$ is uniform.

Graphically, a kernel density plot of the posterior predictive distribution against the data can provide a quick assessment for the model fit.

2.4 Model comparison

To compare the models, we use two measures. The deviance information criterion (DIC) is defined as

$$DIC = \bar{D}(\theta) + \widehat{\text{var}}(D(\theta))/2$$

where $D(\theta) = -2 \log p(\mathbf{y} | \theta)$, $\bar{D}(\theta)$ is the average of $D(\theta)$ over all posterior samples θ and $\widehat{\text{var}}(D(\theta))$ is the sample variance of $D(\theta)$. DIC measures goodness-of-fit and penalizes model complexity; a lower DIC is preferred.

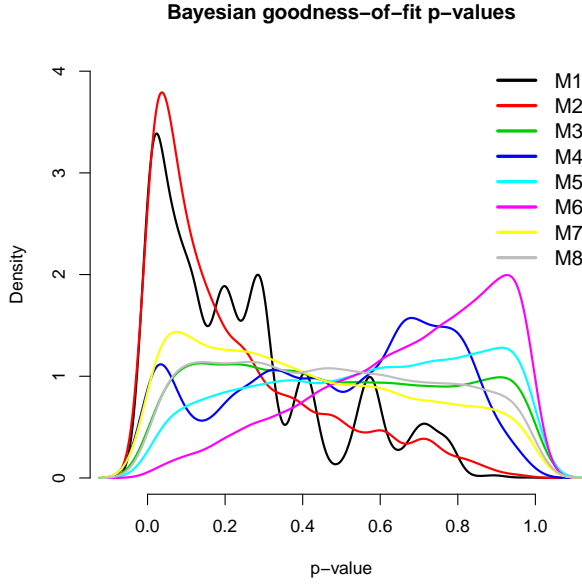


Figure 2: Distribution of p -values for the eight models (identified by color) for the Bayesian goodness-of-fit test.

The second measure is the posterior predictive loss (PPL) which is similar in nature to DIC. It is computed by

$$PPL = G + P = \sum_{i=1}^n (y_i - \mu_i^*)^2 + \sum_{i=1}^n \sigma_i^*$$

where μ_i^* and σ_i^* are the mean and variance of the posterior prediction for observation i . We put more faith in DIC over PPL, since as model complexity grows, PPL is not punished as much as it should.

3. Results

In all subsequent analyses, we retain $B = 10000$ posterior samples after a burn-in period. We first analyze whether the eight models are adequate fits to the data. We obtain a distribution of p -values (based on a χ^2 test) for each model. These are shown in Figure 2. Small p -values give evidence that the model is inadequate. The only two models that seem to be inadequate are M1 and M2, the intercept-only model and the intercept and linear trend model, respectively.

Leave-one-out cross-validation shows similar findings (see Figure 3). Distributions that are uniform give evidence that the model fits. A visual inspection suggests that, again, only M1 and M2 are questionable. However, the K-S test resulted in p -values greater than 0.19 in every case. Hence, there is some evidence that suggests M1 and M2 are inadequate, but no evidence for M3-M8 to be poor fits. Figure 5 shows a plot of observed versus fitted values (on the logit scale) and Figure 4 shows posterior predictions (on the original scale) across time, both for M8.

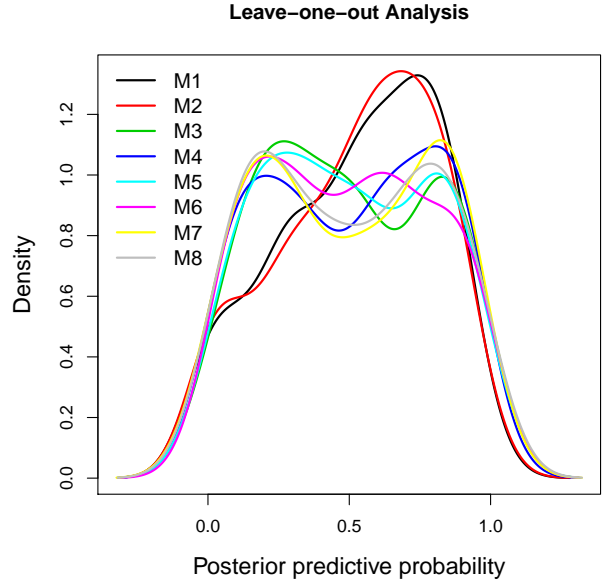


Figure 3: Distributions of the CDF evaluated at each observation based on leave-one-out cross-validation.

Table 2: Model comparison quantities

	DIC	G	P	G+P
M1	-97.47	2.12	2.20	4.32
M2	-96.92	2.10	2.18	4.28
M3	-158.06	1.02	1.14	2.17
M4	-180.62	0.90	0.96	1.87
M5	-158.32	1.00	1.13	2.13
M6	-195.31	0.76	0.82	1.59
M7	-182.75	0.78	0.88	1.67
M8	-199.31	0.65	0.74	1.39

We now address the main questions: is there a time trend? is there an effect due to polling source? and is there an effect due to poll type? We answer these by comparing the eight models. M8 includes all terms (\mathbf{t} , \mathbf{S} , \mathbf{P} as described in 2.1) and is considered the full model. If we remove some of these terms, do we still have a sufficient model?

Table 2 presents the DIC and PPL for each model. M8 does the best under both criteria. M6 (removal of the source effect) is not far behind. This would suggest that there may be no source effect. But removing the time trend (M7) or the polling type (M5) from the full model has a much more dramatic change in DIC, suggesting that both effects are real. In addition, under M8 the posterior distributions for the elements in β corresponding to the source effects all have zero in their 95% HPD (Figure 6).

Given these results, M6 would seem to perform just as well as M8 and perhaps better since it would be less prone to overfitting. This is to say there is a linear time trend, which is steadily increasing in favor of leaving the EU, and that the type of poll has a noticeable impact on

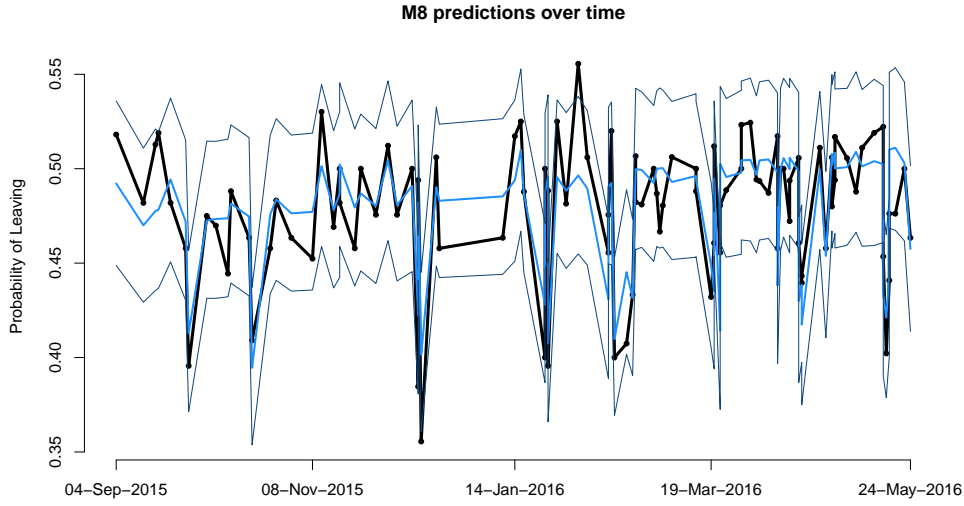


Figure 4: Posterior predictions for M8, back-transformed to the original scale. The black is the data. The middle blue line is the predictive mean and the outer blue lines are equal-tailed 95% posterior intervals.

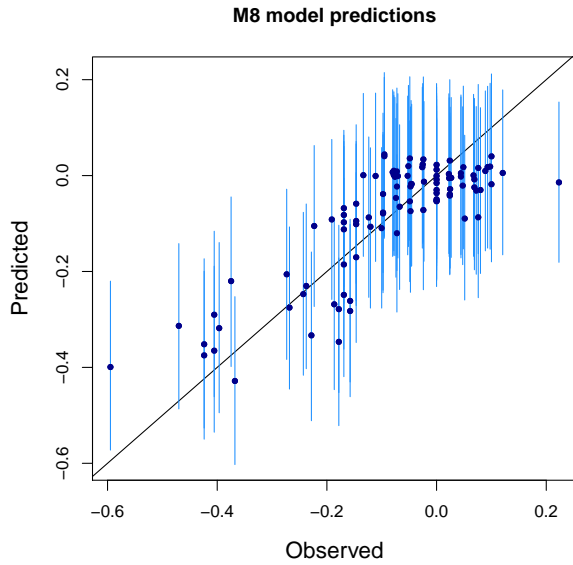


Figure 5: Plot for observed versus fitted for M8. The dots are mean predictions and the lines are equal-tailed 95% predictive intervals. The black line is the line $y = x$.

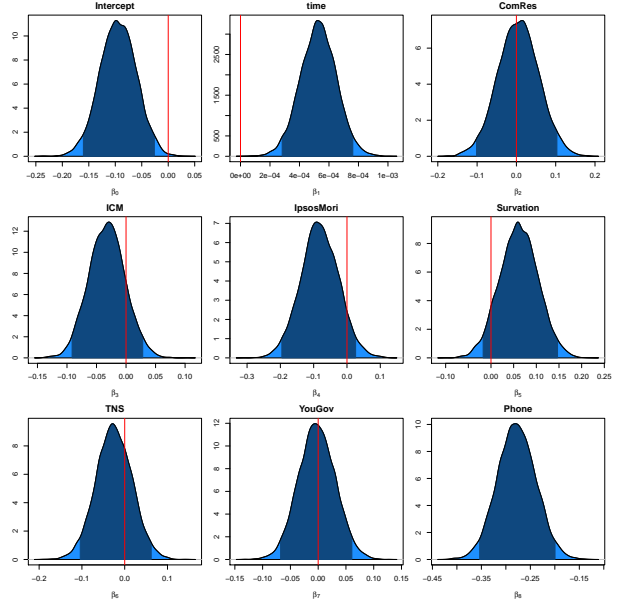


Figure 6: Marginal posterior distributions for β for M8. The shaded regions are the 95% HPD intervals and the red vertical lines mark $x = 0$.

polling results with phone surveys showing preferences toward remaining.

4. Discussion

We have explored the recent UK polls on voter preferences regarding the EU referendum. There is evidence that opinions have changed over time and that we are seeing a steady increase toward leaving the EU. Though, depending on the type of survey (whether online or phone)

we are seeing different polling results. This is likely due in large part because of the accessibility of online surveys and the type of people responding to these polls.

Our regression models provide adequate fits to the data, but some perform better than others. A full model accounting for time, source, and poll type does well. But a model with only time and poll type performs nearly on the same level. In the end, the real voting takes place in a month and many of those people making bets on the outcome are going to lose a lot of money.