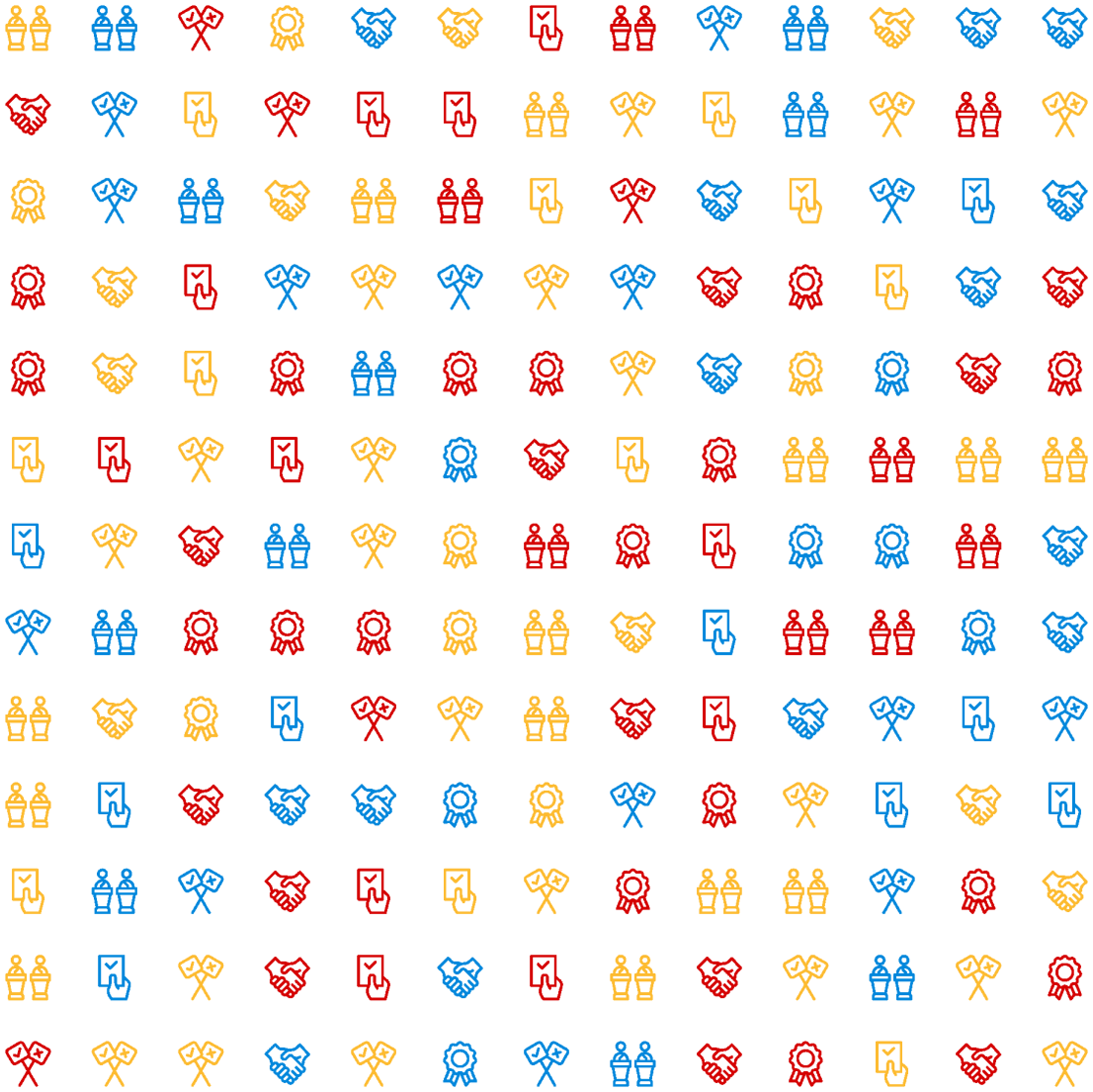
# PollBasePro v0.1.0 (Beta)

# User Guide and Codebook

This version: 11 March 2021



## Contents

## About PollBasePro

PollBasePro is a data package for the `R` programming language. It includes two datasets: `pollbase` and `pollbasepro`. The first provides a long-format and ready-to-analyse version of Mark Pack's (2021) dataset of historic British public opinion polls combined with more recent polling data from Wikipedia. The second provides 24,032 daily estimates of voting intention figures for each of Britain's three largest parties between 26 May 1955 and 11 March 2021.

Both `pollbase` and `pollbasepro` are *living datasets*. Thus, users should endeavour to use only the most recent version of the data when conducting their analyses. This will become more important in the future, as we intend to add new polls and estimates over time. Likewise, we cannot rule out that minor mistakes might have crept into our data processing pipeline. To guard against this, we have made all of our materials available for others to inspect. If you find an error in the code or wish to make a recommendation for a future update, we invite you to raise an issue on the PollBasePro GitHub repository.

## Citing PollBasePro

If you use PollBasePro, you should also cite it. This is good practice and also allows the PollBasePro team to monitor how the data have been used. The project comprises three items: the data, this user guide, and a companion paper. The citations for each item are as follows:

- **Data:** Bailey, J., M. Pack, and L. Mansillo (2021) PollBasePro: Daily Estimates of Aggregate Voting Intention in Great Britain from 1955 to 2021 v.0.1.0 [computer file], March 2021.

- **Documentation:** Bailey, J., M. Pack, and L. Mansillo (2021) PollBasePro v0.1.0: User Guide and Codebook. Retrieved from doi.

- **Paper:** Bailey, J., M. Pack, and L. Mansillo (2021) PollBasePro: Daily Estimates of Aggregate Voting Intention in Great Britain from 1955 to 2021. Retrieved from doi.

## Using PollBasePro

Getting started with PollBasePro in R is simple. It requires only three short steps: first, to install the package; second, to load the package; and, third, to load the data. Once you have taken these three steps, you can then begin analysing the data. Note that PollBasePro is not yet available on CRAN, the service that hosts most R packages. As such, it is not yet possible to install PollBasePro with R's standard `install.packages()` function. Instead, we must install it directly from its GitHub repository. Thankfully this is straightforward and requires only that you run the following code in your R console:

```
# 1. Install the PollBasePro package from GitHub
devtools::install_github("jackobailey/PollBasePro")

# 2. Load the PollBasePro package in R
library(PollBasePro)

# 3. Load the pollbase and pollbasepro datasets
data("pollbase")
data("pollbasepro")
```

Though PollBasePro is first and foremost an R data package, we have sought to accommodate those who use Stata and SPSS too. To make this as easy as possible, we have made both the `pollbase` and `pollbasepro` datasets available as .dta and .sav files. These include all necessary value and variable labels and should work seamlessly with both software packages. You can download the latest version of the data by clicking the links below:

- `pollbase_0.1.0.dta`
- `pollbasepro_0.1.0.dta`
- `pollbase_0.1.0.sav`
- `pollbasepro_0.1.0.sav`


Figure 1: The posterior uncertainty in the estimates that we include in the pollbasepro dataset are correlated with time. This is because polls have become more frequent and have included larger sample sizes as time has passed. Thus, our estimates also become more precise.

Note that the estimates that we include in the `pollbasepro` dataset are *probabilistic*. As such, we include in the dataset both an estimate of the posterior mean of aggregate voting intention for each party on each day *and* the posterior uncertainty in these estimates. As figure 1 makes clear, our uncertainty estimates are not random. Instead, they are correlated with time. This occurs because polls have become more numerous and have tended to include larger sample sizes as time has passed. Thus, our estimates for more recent years are more certain than our estimates for years long far into the past.

We advise all those who use the `pollbasepro` dataset to include this uncertainty in their analyses wherever possible. This is important both because propagating our uncertainty forward is good practice and because such uncertainty serves both to reduce statistical power and to attenuate real and existing relationships in the data. This is possible using "errors-in-variables models." These models work much like regular generalised linear models, though account for measurement error in either the dependent variable, the independent variables, or both. McElreath (2020) provides a good introduction to the intuition behind error-in-variable models. Similarly, Bürkner (2017) provides an easy-to-use interface for fitting such models in `R` using the `brms` package (see also chapter 15.1 in Kurz 2020 for an applied example).

## Technical Details: Estimating Daily Voting Intention

Deriving our estimates is a relatively complex process. As such, it requires much careful consideration. Though we rely on the method put forth in Jackman (2005), there are some issues specific to our case that we must overcome. In this appendix, we elaborate on our choices and build up our modelling process step by step.

To begin, we assume that each poll in the PollBase data, $Poll_i$, takes some value between 0 and 1 and is Normally-distributed around some mean, $\mu_i$, with some known error, $\sigma_i = \sqrt{\frac{Poll_i(1-Poll_i)}{\nu_i}}$, where $\nu_i$ is the sample size of $Poll_i$, $n_i$, divided by the number of days that $Poll_i$ spent in the field, $k_i$.

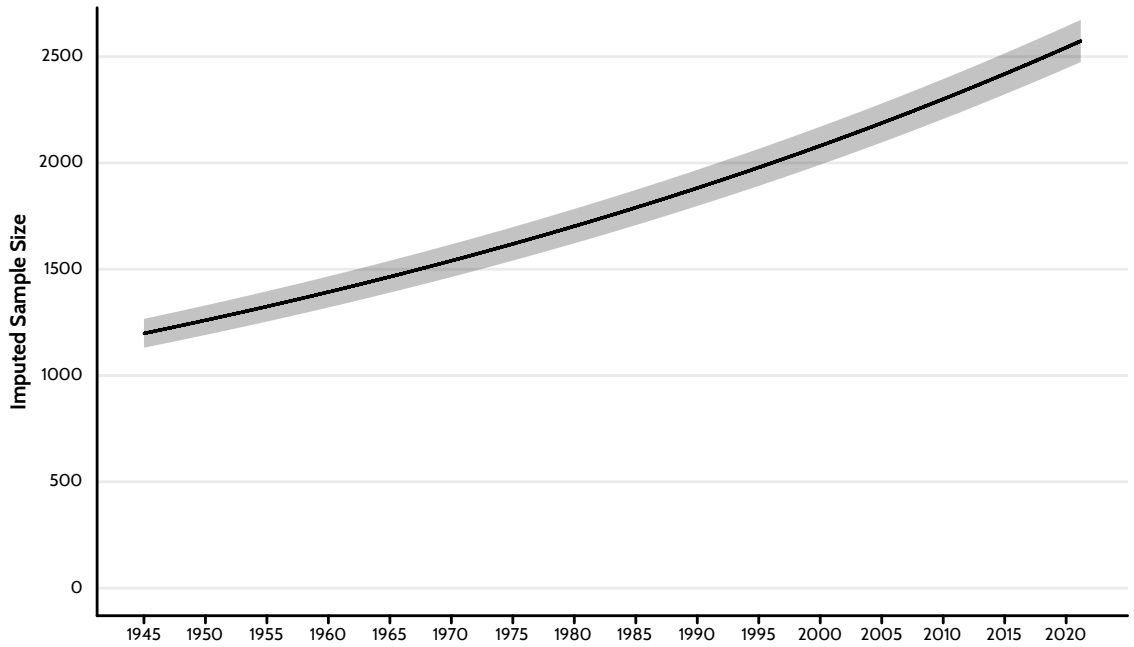



Figure 2: Imputed sample sizes in Britain between 1945 and 2021, estimated using data from Jennings and Wlezien's "Timeline of Elections" dataset (2016).

$$Poll_i \sim \text{Normal}(\mu_i, \sqrt{\sigma^2 + S_i^2}) \quad \text{Likelihood function}$$

$$\mu_i = \alpha_{Day[i]} + \delta_{Pollster[i]} \quad \text{Measurement model on } \mu$$

$$\alpha_t = \alpha_{t-1} + \tau\omega_{t-1} \text{ for } t \text{ in } 2, ..., T-1 \quad \text{Dynamic model on } \alpha_t$$

$$\alpha_T \sim \text{Normal}(\alpha_{T-1}, \tau) \quad \text{Adaptive prior on } \alpha_T$$

$$\delta_j \sim \text{Normal}(0, 0.1) \text{ for } j \text{ in } 1, ..., J \quad \text{Prior on house effects, } \delta$$

$$\omega_t \sim \text{Normal}(0, 0.1) \text{ for } t \text{ in } 1, ..., T-1 \quad \text{Prior on random shocks, } \omega$$

$$\tau \sim \text{Normal}(0, 0.05)^+ \quad \text{Positive prior on scale of innovations, } \tau$$

$$\sigma \sim \text{Exponential}(20) \quad \text{Prior on residual error, } \sigma$$

Unfortunately, the PollBase data that we use to derive our estimates do not include sample sizes. This is a problem, as identifying the model requires that the error in the polls be known. To overcome this problem, we impute likely sample sizes for each poll in the PollBase dataset using similar data from Jennings and Wlezien's (2016) "Timeline of Elections" dataset. Though less comprehensive than PollBase, the Timeline data do include sample sizes.

First, we subset the data to include only polls from the United Kingdom, fit the following Poisson regression model, then use the model to predict likely sample sizes for each day covered in PollBasePro.

We validate the `pollbasepro` data by comparing them to Jennings and Wlezien's "Timeline of Elections" dataset (2016). These data contain 4,302 polls from Britain that ran between 15 June 1943 and 6 June 2017. Given that the `pollbase` data are so comprehensive, it is likely that most of these polls appear in both datasets. Still, the Timeline data provide a good test as they were compiled independently. Our estimates appear well validated. In all cases, correlations between `pollbasepro` and the Timeline data are strong and positive. For the Conservatives, the figure is 90.1% (95% CI: 89.5% to 90.7%); Labour, 92.0% (95% CI: 91.5% to 92.4%); and the Liberals, 91.6% (95% CI: 91.1% to 92.1%).

## Variable List: PollBase (Historical British Election Polls from 1945 to 2021)

| Name | Description |
|---|---|
| `id` | Unique poll identification number |
| `election` | Date of last general election |
| `govt` | Largest party in government after the last general election |
| `start` | First day of fieldwork |
| `end` | Last day of fieldwork |
| `pollster` | Polling company that conducted the poll |
| `n` | Sample size |
| `con` | Voting intention: Conservative |
| `lab` | Voting intention: Labour |
| `lib` | Voting intention: Liberal |
| `con_ldr` | Leader of the Conservative Party |
| `lab_ldr` | Leader of the Labour Party |
| `lib_ldr` | Leader of the Liberals (various forms) |

## Variable List: PollBasePro (Daily British Voting Intention from 1955 to 2021)

| Name | Description |
|---|---|
| `date` | Date |
| `election` | Date of last general election |
| `govt` | Largest party in government after the last general election |
| `con_est` | Posterior mean: Conservative voting intention |
| `con_err` | Posterior error: Conservative voting intention |
| `lab_est` | Posterior mean: Labour voting intention |
| `lab_err` | Posterior error: Labour voting intention |
| `lib_est` | Posterior mean: Liberal voting intention |
| `lib_err` | Posterior error: Liberal voting intention |
| `con_ldr` | Leader of the Conservative Party |
| `lab_ldr` | Leader of the Labour Party |
| `lib_ldr` | Leader of the Liberals (various forms) |
| `week` | Weekly subset indicator |
| `month` | Monthly subset indicator |
| `quarter` | Quarterly subset indicator |
| `year` | Yearly subset indicator |

## Change Log

For the sake of openness and transparency, we provide a change log that lists all updates and changes made to the PollBasePro datasets and documentation over time. If you think that you have found a problem with either, please raise an issue on the project's GitHub repository.

*Version 0.1.0 (Beta)*

- Beta release of data, user guide, and accompanying paper

## References


Bürkner, Paul-Christian. 2017. "Brms: An R Package for Bayesian Multilevel Models Using Stan." *Journal of Statistical Software* 80 (1): 1–28. https://doi.org/10.18637/jss.v080.i01.

Jackman, Simon. 2005. "Pooling the Polls over an Election Campaign." *Australian Journal of Political Science* 40 (4): 499–517. https://doi.org/10.1080/10361140500302472.

Jennings, Will, and Christopher Wlezien. 2016. "The Timeline of Elections: A Comparative Perspective." *American Journal of Political Science* 60 (1): 219–33. https://doi.org/10.1111/ajps.12189.

Kurz, A. Solomon. 2020. "Statistical Rethinking with Brms, Ggplot2, and the Tidyverse: Second Edition (version 0.1.1)." https://bookdown.org/content/4857/.

McElreath, Richard. 2020. *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. Second. CRC Texts in Statistical Science. Boca Raton: CRC Press.

Pack, Mark. 2021. "POLLBASE: OPINION POLLS DATABASE FROM 1943-TODAY." https://www.markpack.org.uk/opinion-polls/.