Revisiting the 2016 US Presidential Election

In the 2016 US presidential election, the Republican candidate Donald Trump surprised many by defeating the Democratic candidate Hillary Clinton. In particular, even right before the election, polls were predicting that Hillary Clinton would win the election by a comfortable margin. Why did preelection polls fail to predict the election outcome? We analyze the polling data, taken from Hufftington post, that include the most recent polls leading up to the election. The dataset we will be analyzing (`polls2016.csv`) has 1395 observations, each representing a different poll, and includes the following variables:

Name	Description
`id`	Poll ID
`state`	U.S. state where poll was fielded
`Clinton`	The poll's estimated level of support for Hillary Clinton
`Trump`	The poll's estimated level of support for Donald Trump
`Undecided`	The poll's estimated percentage of undecided voters
`days_to_election`	Number of days before November 4, 2016.
`electoral_votes`	Number of electoral votes allocated to the state where the poll was fielded (a state-level variable)
`sample_size`	The number of people surveyed in the poll

We will also analyze a dataset (`election2016.csv`) which contains the state-by-state voteshare for each candidate collected from CNN. This data set has the following variables:

Name	Description
`State`	U.S. state where poll was fielded
`Clinton`	The percent of votes Clinton received
`Trump`	The percent of votes Trump received

Question 1

We will begin by calculating the predicted vote share for Hillary Clinton by using the average support rate of the most recent (based on the `days_to_election` variable) polls for each state. If there are multiple polls on the same day, use the average sample size. What is the bias of prediction across states? What is the root mean squared error? Create a histogram of prediction error. Briefly interpret these results.

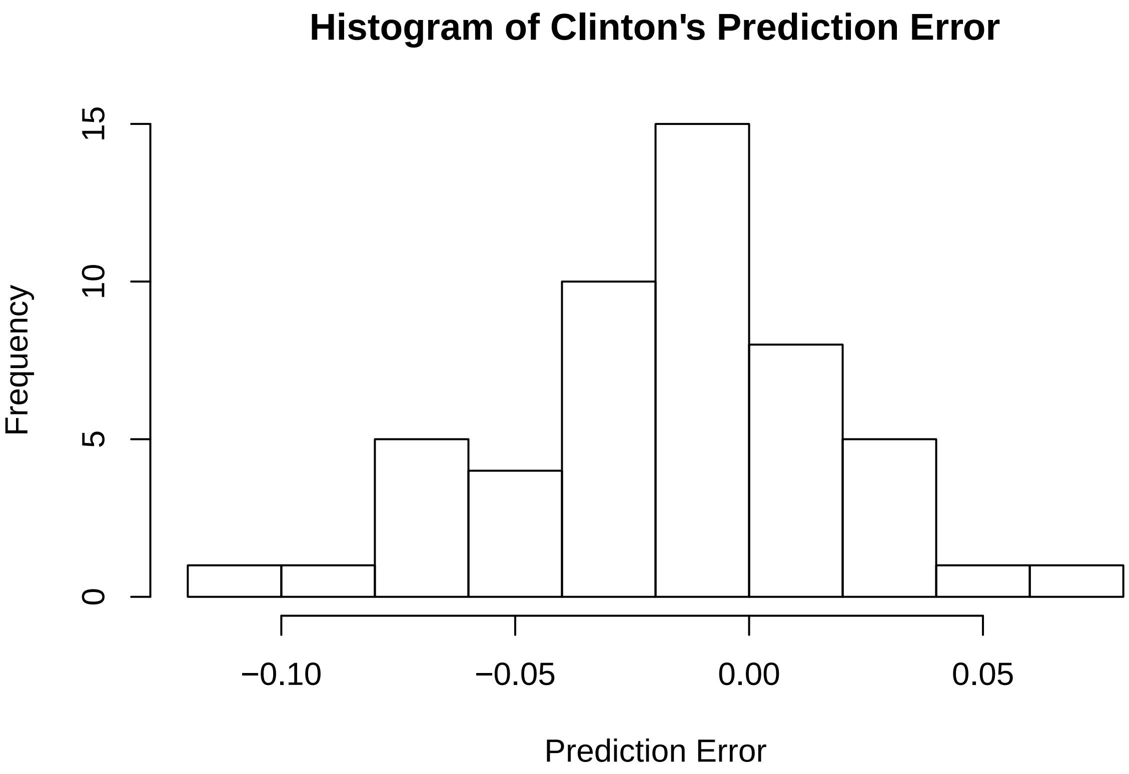
```r
results <- read.csv("data/election2016.csv")
polls <- read.csv("data/polls2016.csv")
state.names <- unique(polls$state)

## Predictions for Clinton
n <- rep(NA, 51)
poll.pred.C <- matrix(NA, nrow = 51, ncol = 3)
row.names(poll.pred.C) <- as.character(state.names)
for (i in 1:51) {
  ## subset the ith state
  state.data <- subset(polls, subset = (state == state.names[i]))
  ## subset the latest polls within the state
  latest <- state.data$days_to_election == min(state.data$days_to_election)
  ## compute the mean of latest polls and store it
  poll.pred.C[i, 1] <- mean(state.data$Clinton[latest])
  n[i] <- mean(state.data$sample_size[latest])
}

## Calculate Bias
Clinton.bias <- poll.pred.C[,1] - results$Clinton
mean(Clinton.bias)
```

```
## [1] -0.01672549
```

```
## Root Mean Squared Error
sqrt(mean((Clinton.bias)^2))
```

```
## [1] 0.03820633
```

```
## Histogram of Bias
hist(Clinton.bias, xlab = "Prediction Error",
     main = "Histogram of Clinton's Prediction Error")
```



The polls under-predicted her voteshare by only 1.67 percentage points. The RMSE is around 3.82 percentage points which tells us there is a substantial amount of variation in the prediction error. Additionally the histogram demonstrates that the prediction error is pretty evenly distributed around 0. In other words, the bias is relatively small.

Question 2

Construct 95% confidence intervals for each of the state-level predictions obtained in the previous question. Plot the prediction against the true result with a 45-degree line to indicate whether the polls under or over predicted Clinton's voteshare. What proportion of the actual election results are contained within these confidence intervals? Does the coverage improve if we correct for the bias of prediction obtained in the previous question? Briefly interpret your results.

Question 3

Repeat the analysis from Questions 1 and 2 for Donald Trump. Compare and interpret your results.

Question 4

We will now explore one hypothesis for Trump's surprising victory in the election: a large proportion of voters whom polls classified as "undecided" cast ballots for Trump on the election day. These voters may not have wanted to admit they supported Trump when answering surveys. It is also possible that they made up their minds right before the election following the FBI announcements. Although we do not have individual data necessary for directly testing this hypothesis, we will predict Trump's electoral college votes under the assumption that all undecided voters voted for Trump. Specifically, run 1000 Monte Carlo simulations under this assumption by computing the probability of winning each state j for Trump as follows:

$P(\text{Trump wins state } j) = P(Z_j > 0.5)$

where Z_j is a Normal random variable with mean $\hat{p}_j$ and standard deviation $\sqrt{\hat{p}_j(1-\hat{p}_j)/n_j}$ with n_j being the sample size of the latest poll for that state and

$\hat{p}_j = \frac{\text{Trump supporters + undecided respondents}}{\text{Trump supporters + Clinton supporters + undecided respondents}}$

Simulate Trump's electoral vote outcomes by sampling its winner using the above probability. In other words, first calculate $\hat{p}_j$ for each state j, then run a simulation where you sample whether Trump wins that state using a draw from a Bernoulli Distribution with the probability of success equal to the above probability $P(\text{Trump wins state } j)$. Present the results using a histogram with a red vertical line representing the actual outcome (Trump = 306). Additionally report the point estimate, standard error, and its 95% confidence interval for the total number of electoral votes for Trump.