Project 1:
Predicting Success in a Presidential Election

Anya Zarembski

September 2022

1 Introduction

The purpose of this project was to investigate if a given candidate will be elected president, based on three factors: height, incumbency status, and the highest office previously served. I decided on this project because I am extremely fascinated by presidential history, especially elections. In fact, studying presidential elections and voting patterns is what initially drew me to statistics. This was a natural project for me to pursue. My guiding question was "What is the relationship between being elected president and height, incumbency, and previous offices held?"

The data used in this project was compiled by myself, containing 125 observations. For all of my data, I used Wikipedia and then compiled it myself into a CSV file. This includes every presidential election from 1796 to 2020. The dataset has 125 observations. It includes candidates from each major party running, and important and influential third party candidates. While there was no strict metric used to determine if a third-party candidate should be included, I generally included any candidates who received electoral college votes, or greater than 10% of the popular vote. For example, in 1992, Ross Perot received almost 19% of the popular vote, so he was included, whereas in 1996, he only received 8%, so he was omitted. The dataset also includes the variables of if a candidate won an election, the year of an election (as an identifier), the name of the candidate (as an identifier), the height of a candidate, what the candidate's highest level of office was at the time of the election, and if the candidate was an incumbent.

1.1 Response Variable

The response variable for this investigation will be a binary win or lose variable. In this dataset, 0 indicates the situation in which a candidate loses, and 1 represents a candidate winning. For the data on who won, I used the *United States presidential election* page. Figure 1 shows the frequency of a candidate winning. We can see that candidates win 46.77% of cases, and lose 53.23%.

It makes sense that there are slightly more losses because of the inclusion of third-party candidates. This data did not require any cleaning.

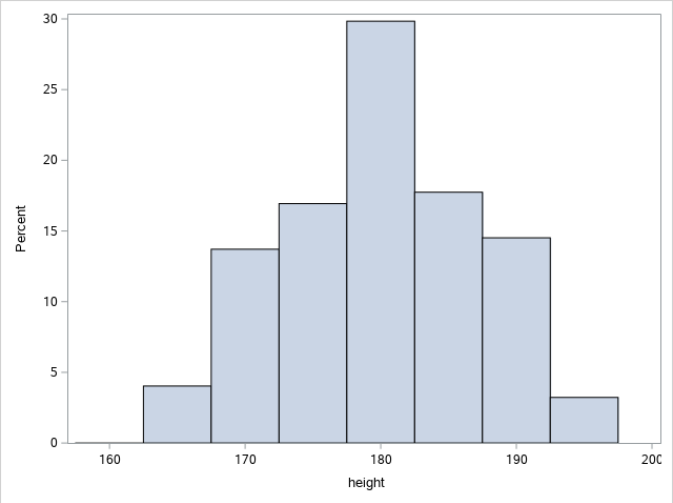
Figure 1: Frequency of Wins in a Presidential Election

win	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	66	53.23	66	53.23
1	58	46.77	124	100.00

1.2 Explanatory Variables

One of the explanatory variables will be height. This is a numeric variable represented in centimeters. I found the heights of each candidate on the *Heights of presidents and presidential candidates of the United States* page on Wikipedia. Based on Figure 2, the heights have a normal distribution, which is to be expected, as height is typically normally distributed. This data did not require any cleaning.

Figure 2: Heights of Presidential Candidates in centimeters



The next explanatory variable is incumbency. A candidate is defined as an incumbent when they are running for an office that they currently are seated in. For example, in 2012, President B. Obama was an incumbent during his campaign for president because he was currently serving as President. This is a binary variable, where 0 represents not being an incumbent, and 1 represents being an incumbent. In Figure 4, it shows that approximately 27% of the

candidates are incumbents. In Figure 5, looking at the column percent shows that 40% of non-incumbents won their elections, while 63% of incumbents won their elections. This would indicate that incumbents are more likely to win a presidential election.

Figure 3: Frequency Table of Incumbency

incumbent	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	91	73.39	91	73.39
1	33	26.61	124	100.00

Figure 4: Contingency Table of Incumbency and Winning

Frequency Percent Row Pct Col Pct	Table of incumbent by win		
	win		
incumbent	0	1	Total
0	54 43.55 59.34 81.82	37 29.84 40.66 63.79	91 73.39
1	12 9.68 36.36 18.18	21 16.94 63.64 36.21	33 26.61
Total	66 53.23	58 46.77	124 100.00

The final explanatory variable is the categorical variable of the highest office served by a candidate at the time of the election. This variable has seven different categories. For the data on the highest office served at the time of the election, I used each candidate's individual Wikipedia pages. The offices used were the following, in order from highest to lowest office:

- President
  - While this may initially seem the same as incumbency status, someone is only counted as an incumbent if they are currently serving as president during an election. If they choose to run for president again, while not currently sitting in office, they would not be an incumbent.
- Vice President

- President's Cabinet
  - This category was difficult to place because while the Secretary of State is clearly a higher office than a Senator or Governor, the Secretary of Transportation would probably not be considered higher by most people. Luckily, this is not an issue for very many candidates, so I simply used my best judgment and considered what they were best known for.
- Governor
- Senator
- Military Officer
  - Once again, this is a difficult office to place. I generally placed being a military officer between the Senate and House for those it concerned, but for candidates who were much better known for their military service as opposed to their political careers, such as President U. S. Grant and President W. H. Harrison, I considered their military service to be their highest office.
- US Representative
- Other
  - This category consists of everything from being an ambassador (Ambassador J. Davis) to no public service at all (President D. Trump).

Figure 5 shows the frequency of the highest offices held by candidates. The most popular office is President, at 28.23%, which is close to our percentage of incumbents, 26.61%, albeit slightly higher due to a few cases of former presidents running again. The next highest office is Governor, 20.97%, followed by Senator, 12.9%. Both of these make sense, as they are extremely high offices that can be held by 50 and 100 people, respectively, at any given time. Vice President may be higher, but there is a smaller pool to draw from, as there can only be one Vice President at any time. The least frequent offices are Other, 3.23%, and a Representative in the House, 6.45%. Both of these also make sense. People not involved in politics rarely run presidential campaigns and it can be difficult for representatives to build a big enough reputation nationally to make it past the primaries as there are so many of them.

Figure 6 can provide more insight. It shows that when the highest office held is President, Vice President, or Military Officer, they win more often than lose. This makes sense. Vice Presidents work alongside the President, so they have closely tied to the Presidency already, and Americans tend to like War Heroes.

Figure 5: Frequency of Highest Office Held by Presidential Candidates

office	Frequency	Percent	Cumulative Frequency	Cumulative Percent
C	15	12.10	15	12.10
G	26	20.97	41	33.06
H	8	6.45	49	39.52
M	9	7.26	58	46.77
O	4	3.23	62	50.00
P	35	28.23	97	78.23
S	16	12.90	113	91.13
VP	11	8.87	124	100.00

Frequency
Percent
Row Pct
Col Pct

Table of office by win

office	win 0	win 1	Total
C	9 7.26 60.00 13.64	6 4.84 40.00 10.34	15 12.10
G	16 12.90 61.54 24.24	10 8.06 38.46 17.24	26 20.97
H	6 4.84 75.00 9.09	2 1.61 25.00 3.45	8 6.45
M	4 3.23 44.44 6.06	5 4.03 55.56 8.62	9 7.26
O	3 2.42 75.00 4.55	1 0.81 25.00 1.72	4 3.23
P	13 10.48 37.14 19.70	22 17.74 62.86 37.93	35 28.23
S	10 8.06 62.50 15.15	6 4.84 37.50 10.34	16 12.90
VP	5 4.03 45.45 7.58	6 4.84 54.55 10.34	11 8.87
Total	66 53.23	58 46.77	124 100.00

Figure 6: Contingency Table between Wins and Highest Office

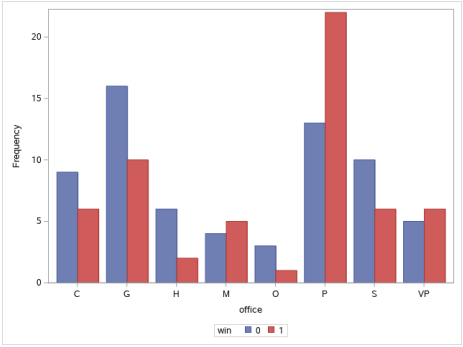


Figure 7: Clustered Bar Chart of Highest Office Frequency, Grouped by Wins

2 Methods

2.1 Specification

Let win_i represent

$$win_i = \begin{cases} 1 & \text{if a candidate wins the presidential election} \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

for all $i \in \{1, 2, ..., 124\}$

Further, let

- H_i = the height of the i^{th} candidate in centimeters
- $I_i = 1$ if the i^{th} candidate is not an incumbent, 0 otherwise
- Vice President, VP, is the baseline highest office held. The other categories the linear predictor are denoted:
  - $C_i = 1$ if the i^{th} candidate's highest office is a position in the President's Cabinet, 0 otherwise
  - $G_i = 1$ if the i^{th} candidate's highest office is Governor, 0 otherwise
  - $HR_i = 1$ if the i^{th} candidate's highest office is US House, 0 otherwise
  - $M_i = 1$ if the i^{th} candidate's highest office is Military Office, 0 otherwise
  - $O_i = 1$ if the i^{th} candidate's highest office falls into the Other category as described in Section 1.2, 0 otherwise
  - $P_i = 1$ if the i^{th} candidate's highest office is President, 0 otherwise
  - $S_i = 1$ if the i^{th} candidate's highest office is Senator, 0 otherwise

The model can be written as:
Random Component:

$$win_i \sim \text{Bernoulli}(\pi_i) \tag{2}$$

where π_i represents the probability of success for the i^{th} candidate.

Logit Link Function:

$$g(\pi_i) = \eta_i = \log(\frac{\pi_i}{1-\pi_i}) \tag{3}$$

Linear Predictor:

$$\begin{aligned}\eta_i = \ & \beta_0 + \beta_1(H_i) + \beta_2(I_i) + \beta_3(C_i) + \beta_4(G_i) + \beta_5(HR_i) \\ & + \beta_6(M_i) + \beta_7(O_i) + \beta_8(P_i) + \beta_9(S_i)\end{aligned} \tag{4}$$

2.2 Justification

A generalized linear model (GLM) is appropriate to model this data because the response variable, win, is binary, and specified in (1). A candidate can either lose or win ($\Omega = \{0, 1\}$). Thus, this response variable is best represented with the Bernoulli distribution, seen in (2). This data does not apply to something like the Normal distribution, because it is discrete, rather than continuous. The Bernoulli distribution takes one parameter, $\pi \in \mathbb{R}$ such that $0 \leq \pi \leq 1$. Finding π using ordinary linear regression ($\pi_i = \beta_0 + \beta_1 x_1 + ... + \beta_n x_n$) could possibly result in $\pi_i < 0$ or $\pi_i > 1$. So, using a generalized linear model helps to link π_i to the explanatory variables to prevent this. We can do this through a function $g()$, for this model the function is specified in (3). Once linked through $g()$, we can create an interpretable linear predictor, like the one specified in (4).

2.3 Explanatory Variables Explained

The model uses three explanatory variables: height, incumbency, and highest office held. Height was chosen because since the advent of everyday photography and especially TV there has been a phenomenon of the taller candidates winning. Incumbency was chosen because incumbents win more often than non-incumbents, as shown in Figure 3.

The highest office held was chosen because it is extremely common for Vice Presidents to become presidents, as well as that it is actually very uncommon for sitting Congressmen to be elected presidents. Only one sitting member of the House of Representatives (President J. Garfield) and three sitting Senators (Presidents W. Harding, J. Kennedy, and B. Obama) have been elected to the presidency. I believed there would be some patterns in what kind of offices Americans trust with the presidency.

3 Results

3.1 Complete Separation

When creating the model, everything went smoothly. There was no evidence of complete separation, as none of the variables had any categories that were perfect predictors. This can be seen in Figures 4 and 6, where none of the categories in either categorical variable (incumbency or office) have 0 when crossed with wins, so there are no perfect predictors.

3.2 The Model

The linear predictor can be written as:

$$\eta_i = -5.384 + 0.034(H_i) - 0.507(I_i) - 0.514(C_i) - 0.536(G_i) - 1.295(HR_i)$$
$$+0.193(M_i) - 1.377(O_i) - 0.108(P_i) - 0.652(S_i)$$

where the probability of success can be defined as:

$$\pi_i = \frac{e^{\eta_i}}{1 + e^{\eta_i}}$$

The coefficients for the model were determined through maximum likelihood estimates.

3.3 Interpretation

With these equations, as a result, we are able to say the following: Holding all other explanatory variables constant,

1. for every centimeter increase in height, the odds of winning a Presidential Election change by a factor of $e^{0.034}$.

2. the odds of winning a Presidential Election change by a factor of $e^{-0.507}$ when a candidate is not an incumbent, as opposed to being an incumbent.

3. the odds of winning a Presidential Election change by a factor of $e^{-0.108}$ when a candidate's highest office served is President, as opposed to the highest office being Vice President.

We can interpret every category in the office variable in a fashion similar to 3. 3 is just one example of the 28 combinations of offices to compare. We can also create confidence intervals to better understand how each variable changes the odds.
For example, for 1, we can say: Holding all other explanatory variables constant, we are 95% confident that for every centimeter increase in height, the odds of winning a Presidential Election change by a factor between $e^{-0.0174}$ and $e^{0.0863}$.
For number 2 we can say: Holding all other explanatory variables constant, we are 95% confident that the odds of winning a Presidential Election change by a factor between $e^{-3.7991}$ and $e^{2.7839}$ when a candidate is not an incumbent, as opposed to being an incumbent.
For 3 we can say: Holding all other explanatory variables constant, we are 95% confident the odds of winning a Presidential Election change by a factor between $e^{-3.5113}$ and $e^{3.293}$ when a candidate's highest office served is President, as opposed to the highest office being Vice President.

3.4 Hypothesis Testing

3.4.1 Entire Model Significance

$H_0 : \beta_0 = \beta_1 = \beta_2 = ... = \beta_9 = 0$
H_a : At least one$\beta \neq 0$
Maximum Likelihood Test Statistic: 9.9822
Null Distribution: $\chi^2(9)$
P-value: P($\chi^2(9)$ ¿ 9.9822) = 0.3519
P-value = 0.3519 $> \alpha = .05$. We fail to reject H_0. We do not believe that our model is a significant predictor of if a candidate will win a presidential election.

3.4.2 Incumbent Significance

$H_0 : \beta_2 = 0$
$H_a : \beta_2 \neq 0$
Maximum Likelihood Test Statistic: 0.1204
Null Distribution: $\chi^2(1)$
P-value: P($\chi^2(1)$ ¿ 0.1204) = 0.7286
P-value = 0.7286 $> \alpha = .05$. We fail to reject H_0. We do not believe that incumbency is a significant predictor of if a candidate will win a presidential election.

3.5 A Prediction for 2024

There are many rumored candidates for the 2024 Presidential Election, including former President Trump, President Biden, Stacy Abrams, and Liz Cheney. We can make predictions for these potential candidates and their potential success in 2024. Let's start with Biden:

Joe Biden is 182 cm tall, will be the incumbent, and the highest office he will have served in will be the presidency ($\mathrm{H}_{\mathrm{Biden}} = 182$, $I_{\mathrm{Biden}} = 0$, $P_{\mathrm{Biden}} = 1$, all other variables = 0). We can plug these numbers in to arrive at:

$$\pi_{\mathrm{Biden}} = \frac{e^{-5.3841+0.0335(182)+-0.5070(1)+-0.1076(1)}}{1+e^{-5.3841+0.0335(182)+-0.5070(1)+-0.1076(1)}} = .647$$

For Donald Trump, $H_{\mathrm{Trump}} = 191$, $I_{\mathrm{Trump}} = 1$, $P_{\mathrm{Trump}} = 1$, all other variables = 0). So $\pi_{\mathrm{Trump}} = 0.599$. For Stacy Abrams, $H_{\mathrm{Abrams}} = 160$, $I_{\mathrm{Abrams}} = 1$, $O_{\mathrm{Abrams}} = 1$, all other variables = 0). So $\pi_{\mathrm{Abrams}} = 0.138$. For Liz Cheney, $H_{\mathrm{Cheney}} = 160$, $I_{\mathrm{Cheney}} = 1$, $HR_{\mathrm{Cheney}} = 1$, all other variables = 0). So $\pi_{\mathrm{Cheney}} = 0.148$.

This gives us a very interesting result. Abrams and Cheney only have one variable different between the two: office, but it resulted in a 1% difference in their probability of getting elected. Abrams is currently running for governor of Georgia; let's see how π_{Abrams} changes if she wins her governor race, then runs for president. If Abrams's highest office was governor $\pi_{\mathrm{Abrams}} = .2701$. This is a drastic change!

4 Conclusions

In the model above, we were not able to prove that our model or its variables were significant in predicting if a given candidate will win the presidential election. The most drastic change in odds comes from changing the level of the highest office, as seen in the example with Stacy Abrams.

For further research into this topic, I would love to explore other explanatory variables. I considered a binary variable denoting if the candidate was descended from a major political figure, and I think this could be interesting to explore. It would certainly change Liz Cheney's odds, as her father is former Vice President Dick Cheney. I also think including a gender variable would be interesting, however, right now there is only one female who has ever run and made it far enough to appear on this dataset, Hillary Clinton in 2016. This would obviously lead to complete separation. It could be remedied by adding Vice Presidential Candidates because then you have one example of a woman winning and three losings. The question would then become "What is the relationship between these explanatory variables and being elected President OR Vice President?"

I think it would also be beneficial to use the difference between the two candidates' height, rather than their height generally as a variable. This would make it more clear because two tall candidates could be running at once, so overall they score well, but really it doesn't matter because they look the same height.

5 Appendix

```
/*Name: Anya Zarembski
Project: Stat172 Project 1
Date: 9/30/2022*/

/*Read in dataset*/
proc import out = presidents
datafile = "/home/u57861990/Data/presidents_w_l.csv"
dbms = csv replace;
guessingrows = 125;
run;

/*Explore win variable
Produce a frequency table of winning

We find that 46.77% of candidates won*/
proc freq data = presidents;
tables win;
run;

/*Explore height variable
Produce a histogram of height with a bin size of 5

We find that height is normally distributed*/
proc sgplot data = presidents;
histogram height /
BINWIDTH = 5;
run;

/*Explore incumbent variable
Produce frequency table of incumbent;
Produce a contingency table for win and incumbent

We find that 26.61% of candidates are incumbents and
incumbents win their elections
63% of the time while non-incumbents only win 40%*/
proc freq data = presidents;
tables incumbent;
run;

proc freq data = presidents;
tables incumbent*win;
run;
```

```
/*Explore Office variable
Produce a bar graph (grouped by win) and
frequency table of office frequency;
Produce a contingency table of office and win

We find that 28.23% of candidates have served as president.
We find that only 3.23% of candidates have
not served in one of the other categories
We find that candidates whose highest
office is president win 62.86% of the time,
followed by Military Office, who win
55.56% of the time.
Candiates whose highest office is
Miltary Office, Presidents, and Vice Presidents
win more than lose, but candidates whose
highest office is Senate, Governor, House,
Cabinet, and Other lose more than win.
*/

proc freq data = presidents;
tables office;
run;

proc freq data = presidents;
tables office*win;
run;

proc sgplot data = presidents;
vbar office /
group = win
groupdisplay = cluster;
run;

/*This is where we build our model.
We classify incumbent and office as categorical
and set our event = to winning a presidential election
We set clparm = both to get maximum liklihood and wald*/

proc logistic data = presidents;
class incumbent office/ param = reference;
model win(event = '1') = height incumbent office/clparm = both;
run;
```