Data Analytics
Automatics and Robotics II-cycle 2023/2024

Predicting presidential elections in United States of America based on elections between 1992-2020 & gun ownership, Human Development Index and unemployment data.

Contributors: Eng. Jakub Szczypek, Eng. Michał Rola

Project supervisors: Prof. Jerzy Baranowski, MEng Adrian Dudek

Table of Contents

1 Problem formulation

Project focused on creating two models and comparing them. Models were tasked with predicting results of presidential elections in United States of America based on gun ownership percentage, Human Development Index, and unemployment percentage of each state. The idea was to create a model that would be able to predict future results of presidential elections in USA.

1.1 Data

Data came from several sources. Below are links to where the data was taken from and examples of how the data looked like before transcription:

- https://www.270towin.com/states/ – Page provides data on how each state voted in presidential elections. The data goes as far as first presidential elections of United States in 1789. One of the graphs, which can be found on the site is shown in Figure 1.

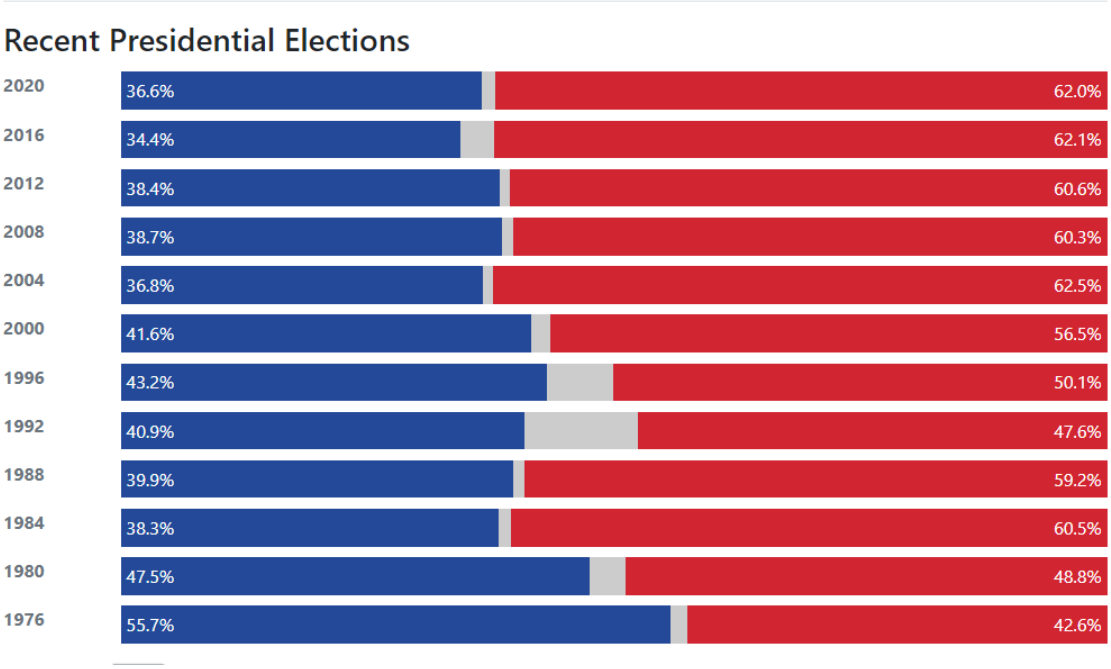

Figure 1. Percentile support for parties (Blue/Left – Democrats, Red/Right – Republicans) over the years in Alabama

- https://www.rand.org/research/gun-policy/gun-ownership.html – A page containing data on gun ownership in each US state from 1980 to 2020. The data was available for download in the form of a ".xlsx" file. A table context is shown in Figure 2. HFR column contains gun ownership in 0-1 scale, those values were multiplied by 100 to obtain 0-100 scale.

FIP	Year	STATE	HFR	HFR_se	universl	permit	Fem_FS_S	Male_FS_S	BRFSS	GALLUP	GSS	PEW	HuntLic	GunsAmm	BackChk	PewQChng	BS1	BS2	BS3
1	1980	Alabama	0,608	0,031	0	0	0,824324324	0,833795	-9	0,55395	0,583632	-9	0,291102	-0,50916	-9	0	0	0	0
1	1981	Alabama	0,597	0,047	0	0	0,692307692	0,831126	-9	-9	-9	-9	0,294962	-0,61895	-9	0	1	0	0
1	1982	Alabama	0,661	0,036	0	0	0,77173913	0,821429	-9	-9	0,655196	-9	0,290545	-0,52669	-9	0	2	0	0
1	1983	Alabama	0,586	0,038	0	0	0,688172043	0,819277	-9	0,61144	-9	-9	0,284983	-0,71323	-9	0	3	0	0
1	1984	Alabama	0,624	0,036	0	0	0,71	0,775956	-9	-9	0,626933	-9	0,281622	-0,7333	-9	0	4	0	0
1	1985	Alabama	0,644	0,031	0	0	0,755555556	0,835294	-9	0,611974	0,662106	-9	0,278214	-0,71912	-9	0	5	0	0
1	1986	Alabama	0,567	0,038	0	0	0,686868687	0,777778	-9	0,596843	-9	-9	0,275302	-0,73223	-9	0	6	0	0
1	1987	Alabama	0,609	0,036	0	0	0,711340206	0,795455	-9	-9	0,586605	-9	0,278852	-0,84291	-9	0	7	0	0
1	1988	Alabama	0,606	0,031	0	0	0,638095238	0,804071	-9	0,768345	0,505046	-9	0,273267	-0,83557	-9	0	7,91	0,09	0
1	1989	Alabama	0,627	0,031	0	0	0,714285714	0,801471	-9	0,702464	0,583599	-9	0,269243	-0,64693	-9	0	8,73	0,27	0
1	1990	Alabama	0,641	0,031	0	0	0,767676768	0,860241	-9	0,652096	0,620018	-9	0,255213	-0,75243	-9	0	9,45	0,55	0
1	1991	Alabama	0,632	0,031	0	0	0,65625	0,814898	-9	0,699918	0,616141	-9	0,24921	-0,80607	-9	0	10,09	0,91	0
1	1992	Alabama	0,552	0,047	0	0	0,717391304	0,765661	-9	-9	-9	-9	0,248809	-0,82324	-9	0	10,64	1,36	0
1	1993	Alabama	0,6	0,031	0	0	0,71559633	0,794118	-9	0,682881	0,541685	-9	0,247217	-0,63105	-9	0	11,09	1,91	0
1	1994	Alabama	0,571	0,036	0	0	0,657894737	0,77512	-9	-9	0,545011	-9	0,244709	-0,61137	-9	0	11,45	2,55	0
1	1995	Alabama	0,554	0,047	0	0	0,639534884	0,786164	-9	-9	-9	-9	0,250194	-0,32559	-9	0	11,73	3,27	0
1	1996	Alabama	0,568	0,031	0	0	0,699029126	0,8	-9	0,502761	0,569423	-9	0,247113	-0,12192	-9	0	11,91	4,09	0
1	1997	Alabama	0,557	0,029	0	0	0,697247706	0,781638	-9	0,594295	-9	0,57044	0,245158	-0,15224	-9	0	12	5	0
1	1998	Alabama	0,577	0,036	0	0	0,657657658	0,79476	-9	-9	0,546264	-9	0,243714	0,004365	-9	0	12	6	0
1	1999	Alabama	0,537	0,036	0	0	0,697916667	0,801743	-9	0,520277	-9	-9	0,240078	-0,28092	0,977565	0	12	7	0
1	2000	Alabama	0,538	0,026	0	0	0,616071429	0,760085	-9	0,57566	0,435643	0,606166	0,243389	-0,15688	0,849506	0	12	7,91	0,09
1	2001	Alabama	0,522	0,021	0	0	0,608247423	0,790361	0,519736	-9	-9	-9	0,244353	-0,15625	0,882527	0	12	8,73	0,27
1	2002	Alabama	0,553	0,02	0	0	0,565656566	0,783133	0,575971	-9	0,440898	-9	0,240564	-0,16242	0,95191	0	12	9,45	0,55
1	2003	Alabama	0,504	0,032	0	0	0,697916667	0,781176	-9	-9	-9	0,495892	0,237966	-0,19351	0,971231	0	12	10,09	0,91
1	2004	Alabama	0,516	0,018	0	0	0,577319588	0,718468	0,525519	-9	0,504182	0,509307	0,239216	-0,19682	0,95049	0	12	10,64	1,36
1	2005	Alabama	0,485	0,044	0	0	0,645833333	0,699317	-9	-9	-9	-9	0,236233	-0,23378	0,928751	0	12	11,09	1,91
1	2006	Alabama	0,468	0,035	0	0	0,572916667	0,71281	-9	-9	0,418362	-9	0,235138	-0,20556	0,840486	0	12	11,45	2,55
1	2007	Alabama	0,481	0,032	0	0	0,56557377	0,695745	-9	-9	-9	0,517509	0,228387	-0,21196	0,87247	0	12	11,73	3,27

Figure 2. Sample of data about gun ownership in each state

Varname	Description
FIP	State FIP value
Year	Year
STATE	State name
HFR	Factor scores for household firearm ownership latent factor
HFR_se	Standard errors of factor scores for household firearm ownership latent factor
universl	State has universal background checks law (1=yes; 0=no)
permit	State has permit to purchase law (1=yes; 0=no)
Fem_FS_S	Female Firearm Suicide/Total Male Suicide *100
Male_FS_S	Male Firearm Suicide/Total Male Suicide *100
BRFSS	BRFSS state-level survey estimate
GALLUP	Gallup state-level survey estimate
GSS	General Social Survey state-level survey estimate
PEW	Pew state-level survey estimate
HuntLic	Square root of Hunting Licenses/Population
GunsAmmo	Within-year standardization of the square root of Guns and Ammo Subscriptions/Population
BackChk	Within-year standardization of Background Checks (without checks for permits)/Population
PewQChng	Indicator for Pew firearm ownership question wording change
BS1	Blended linear spline 1 represents roughly 1980–1992
BS2	Blended linear spline 2 represents roughly 1993–2004
BS3	Blended linear spline 3 represents roughly 2005–2016

Figure 3. Description of gun ownership database

- https://globaldatalab.org/shdi/table/shdi/USA/?levels=1+4&interpolation=0&extrapolation=0 - The database on HDI around the World from 1990 to 2021. This data could also be downloaded as a ".xlsx" file. A table context is shown in Figure 4.

Country	Continent	ISO_Code	Level	GDLCODE	Region	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005
United States	America	USA	National	USAt	Total	0,872	0,873	0,877	0,88	0,884	0,885	0,887	0,889	0,893	0,889	0,891	0,892	0,893	0,895	0,898	0,9
United States	America	USA	Subnat	USAr101	Alabama	0,834	0,835	0,839	0,843	0,846	0,847	0,85	0,852	0,855	0,851	0,853	0,854	0,855	0,857	0,861	0,863
United States	America	USA	Subnat	USAr102	Alaska	0,912	0,912	0,915	0,917	0,92	0,92	0,921	0,922	0,921	0,914	0,913	0,915	0,916	0,916	0,919	0,921
United States	America	USA	Subnat	USAr103	Arizona	0,873	0,873	0,876	0,879	0,882	0,882	0,884	0,885	0,89	0,886	0,888	0,888	0,888	0,89	0,893	0,895
United States	America	USA	Subnat	USAr104	Arkansas	0,833	0,835	0,839	0,842	0,846	0,847	0,85	0,851	0,854	0,851	0,853	0,853	0,855	0,857	0,861	0,862
United States	America	USA	Subnat	USAr105	California	0,879	0,879	0,883	0,886	0,889	0,889	0,891	0,893	0,897	0,893	0,897	0,897	0,898	0,9	0,905	0,907
United States	America	USA	Subnat	USAr106	Colorado	0,903	0,903	0,907	0,91	0,912	0,913	0,915	0,916	0,922	0,918	0,921	0,921	0,92	0,92	0,922	0,923
United States	America	USA	Subnat	USAr107	Connecticı	0,906	0,907	0,911	0,914	0,917	0,918	0,92	0,922	0,926	0,921	0,925	0,927	0,927	0,928	0,934	0,936
United States	America	USA	Subnat	USAr108	Delaware	0,89	0,892	0,896	0,9	0,903	0,905	0,907	0,909	0,915	0,912	0,913	0,913	0,913	0,914	0,915	0,915
United States	America	USA	Subnat	USAr109	District of	0,867	0,87	0,874	0,877	0,88	0,88	0,885	0,888	0,894	0,891	0,896	0,897	0,899	0,901	0,905	0,906

Figure 4. Sample of Human Development Index database

- https://www.ncsl.org/labor-and-employment/state-unemployment-rates – Database on state-individual unemployment in the USA. Database provides monthly data so there was a need to calculate an average for each year. The database was available for download as ".xlsx" file and is shown in Figure 5. The Unemployment/Rate column had necessary data.

States and selected areas: Employment status of the civilian noninstitutional population, January 1976 to date, seasonally adjusted

FIPS Code	State and area	Period		Civilian non-institutional population	Civilian labor force					
					Total	Percent of population	Employment		Unemployment	
		Year	Month				Total	Percent of population	Total	Rate
01	Alabama	1976	01	2 605 000	1 484 555	57,0	1 386 023	53,2	98 532	6,6
02	Alaska	1976	01	232 000	160 183	69,0	148 820	64,1	11 363	7,1
04	Arizona	1976	01	1 621 000	964 120	59,5	865 871	53,4	98 249	10,2
05	Arkansas	1976	01	1 536 000	889 044	57,9	824 395	53,7	64 649	7,3
06	California	1976	01	15 621 000	9 774 280	62,6	8 875 685	56,8	898 595	9,2
037	Los Angeles County	1976	01	5 273 000	3 381 856	64,1	3 081 806	58,4	300 050	8,9
08	Colorado	1976	01	1 832 000	1 230 966	67,2	1 160 104	63,3	70 862	5,8
09	Connecticut	1976	01	2 248 000	1 442 847	64,2	1 301 974	57,9	140 873	9,8
10	Delaware	1976	01	417 000	261 418	62,7	240 543	57,7	20 875	8,0
11	District of Columbia	1976	01	520 000	334 691	64,4	305 677	58,8	29 014	8,7
12	Florida	1976	01	6 421 000	3 584 876	55,8	3 238 460	50,4	346 416	9,7
13	Georgia	1976	01	3 522 000	2 244 770	63,7	2 056 467	58,4	188 303	8,4
15	Hawaii	1976	01	583 000	407 457	69,9	367 586	63,1	39 871	9,8
16	Idaho	1976	01	580 000	365 606	63,0	345 182	59,5	20 424	5,6
17	Illinois	1976	01	8 127 000	5 123 609	63,0	4 786 415	58,9	337 194	6,6
18	Indiana	1976	01	3 789 000	2 425 346	64,0	2 265 849	59,8	159 497	6,6
19	Iowa	1976	01	2 066 000	1 329 798	64,4	1 272 281	61,6	57 517	4,3
20	Kansas	1976	01	1 635 000	1 065 059	65,1	1 020 396	62,4	44 663	4,2
21	Kentucky	1976	01	2 459 000	1 486 550	60,5	1 402 625	57,0	83 925	5,6
22	Louisiana	1976	01	2 672 000	1 517 348	56,8	1 421 440	53,2	95 908	6,3
23	Maine	1976	01	764 000	474 371	62,1	433 260	56,7	41 111	8,7

Figure 5. Sample of unemployment database for each USA state

Since 1992 was the earliest common year for which all data was easily accessible, and the last US elections took place in 2020, that was the timeframe used for creating the model. This period was also picked, because there was no major turmoil inside the country during this time.

1.2 Directed Acyclic Graph

Predictors picked for the model are as follows:

- Gun ownership percentage (G)
- Human Development Index (HDI)
- Unemployment percentage (U)

Unemployment was a pipe (confounder) because it influenced both Human Development Index (Per Capita Income is one of HDI's parameters) and results of presidential elections in US (V).

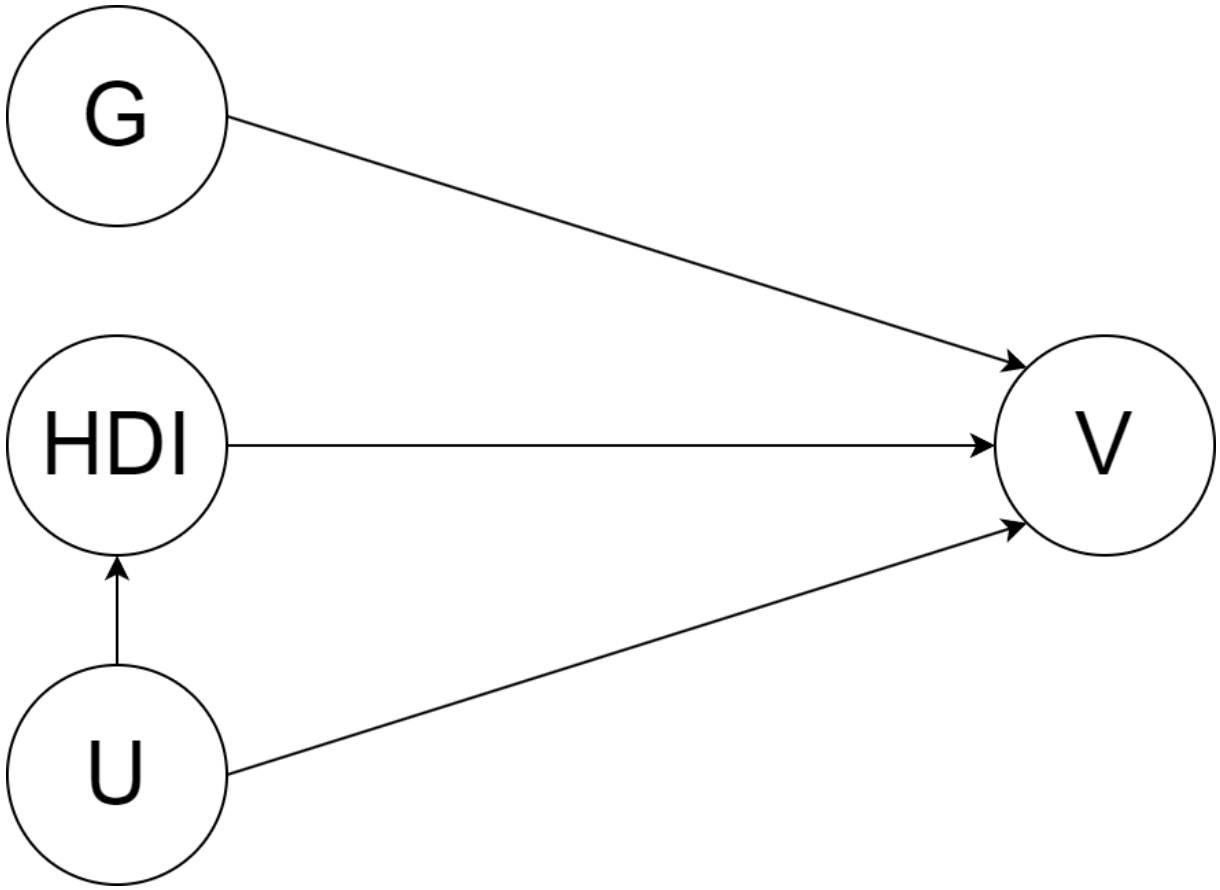

Figure 6. DAG of suspected dependency between predictors and results

2 Data preprocessing

Since data was not presented immediately in a single file then it was required to create a separate target database. It was decided that data would be stored in a ".xlsx" file and contain following columns: "Year", "State", "Democratic %", "Gun ownership %", "HDI" and "Unemployment %". These columns were to be filled with data from the databases mentioned in Chapter 1.

The Year and State columns were created manually. In the first 51 cells 1992 was entered and in the following rows formula was used to add 4 to the value 51 rows above the cell, until getting to 2020. Afterwards, formulas were converted to values, so there wouldn't be any issues after sorting/filtering the data. States were inputted into first 51 rows and then copied and pasted.

The data in "Democratic %" column had to be transcribed manually from the page provided in Chapter 1. No other database could be found that could be downloaded as a ".xlsx" or ".csv" file.

Data in "Gun ownership %" and "HDI" columns was copied.

To get the "Unemployment %" data, it was beneficial to create a script that would calculate necessary information. Following steps were taken to obtain this data:

1. Firstly, columns containing years, months, state and unemployment percentage were selected. This data was obtained from the given database in Chapter 1.
2. Next, average values of unemployment for each state over the years were calculated. The results were saved to a new ".xlsx" file.
3. Lastly, obtained data was then copied directly into main database.

Main database is shown in Figure 7.

Year	State	Democratic %	Gun ownership %	HDI	Unemployment %
1992	Alabama	40.9	55.2	0.839	7.525
1996	Alabama	43.2	56.8	0.85	5.208
2000	Alabama	41.6	53.8	0.853	4.625
2004	Alabama	36.8	51.6	0.861	5.575
2008	Alabama	38.7	50.3	0.867	5.883
2012	Alabama	38.4	54.3	0.876	8.158
2016	Alabama	34.4	52.8	0.882	5.908
2020	Alabama	36.6	55.5	0.879	6.408
1992	Alaska	30.3	68.5	0.915	8.858
1996	Alaska	33.3	66.3	0.921	7.425
2000	Alaska	27.7	60.1	0.913	6.317
2004	Alaska	35.5	58	0.919	7.417
2008	Alaska	37.9	61.3	0.93	6.475
2012	Alaska	40.8	60.7	0.934	7.25
2016	Alaska	36.6	57.2	0.935	6.6
2020	Alaska	42.8	64.5	0.931	8.35
1992	Arizona	36.5	54.7	0.876	7.567
1996	Arizona	46.5	44.9	0.884	5.492
2000	Arizona	44.7	42.7	0.888	3.983
2004	Arizona	44.4	36.8	0.893	4.967
2008	Arizona	45.1	38.2	0.901	5.792
2012	Arizona	44.6	41.5	0.906	8.35
2016	Arizona	45.1	36	0.909	5.483
2020	Arizona	49.4	46.3	0.907	7.825
1992	Arkansas	53.2	53.3	0.839	7
1996	Arkansas	53.7	58	0.85	5.167
2000	Arkansas	45.9	54.2	0.853	4.175
2004	Arkansas	44.6	54.4	0.861	5.583
2008	Arkansas	38.9	50.8	0.867	5.442
2012	Arkansas	36.9	54.4	0.876	7.267
2016	Arkansas	33.7	51.8	0.882	3.983
2020	Arkansas	34.8	57.2	0.879	6.183

Figure 7. Sample of main database

3 Model

It was decided that project would focus on comparison between linear regression model and polynomial regression model to check, which would be more accurate at predicting results of presidential elections in US.

Both models used normal distribution (1).

$$V \sim N(\mu, \sigma) \tag{1}$$

3.1 Data

Since focus of the project was to check, which model is better, both models were fed same data:

- Results of presidential elections in United States between 1992 and 2020 (V),
- Gun ownership percentage (G),
- Human Development Index (HDI),
- Unemployment percentage (U).

3.2 Linear regression model

Linear regression is a sum of bias and predictors multiplied by parameters (2). Those models are simpler to implement, but they can only predict linearly dependent data.

$$\mu = \alpha + \beta_G * G + \beta_{HDI} * HDI + \beta_U * U \tag{2}$$

3.3 Polynomial regression model

Polynomial regression is a sum of bias and subsequent powers of predictors multiplied by parameters (3). Thanks to this it is possible to obtain bigger fluctuations, so small change in predictors values can cause big change in results. It is both a good and a bad thing. Polynomial regression models can predict more complex data, but they are less stable and harder to implement.

$$\mu = \alpha + \beta_{G1} * G + \beta_{G2} * G^2 + \beta_{HDI1} * HDI + \beta_{HDI2} * HDI^2 + \beta_{U1} * U + \beta_{U2} * U^2 \tag{3}$$

4 Priors

4.1 Linear regression model

First of all, correlation matrix for data was created to check how gun ownership percentage, Human Development Index and unemployment percentage correlate with votes for Democratic candidates. This information was highlighted in Figure 8.

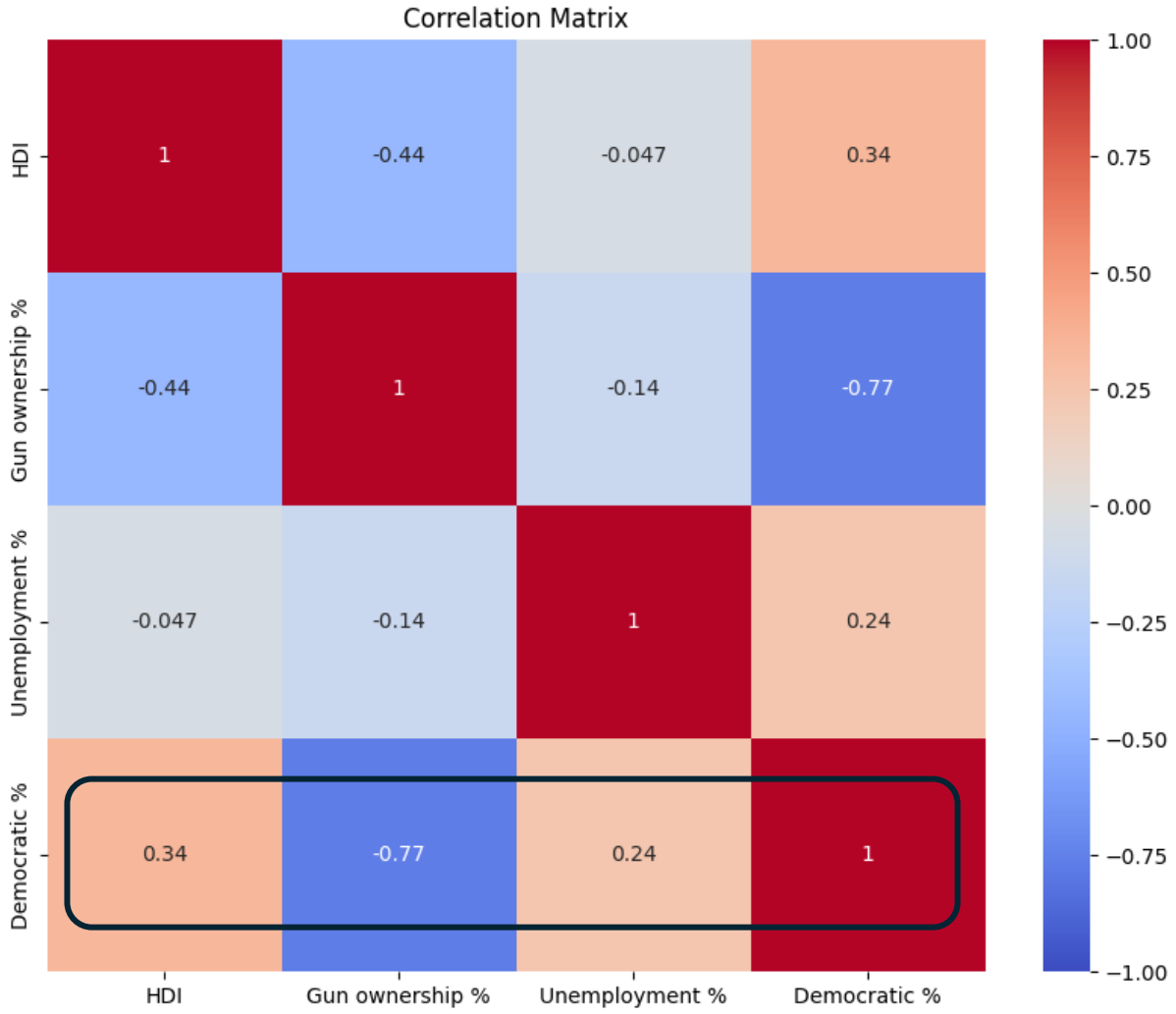

Figure 8. Correlation matrix

Something that was noticed, gun ownership percentage has a strong negative correlation with percentage of votes for Democratic candidates. When it comes to Human Development Index and unemployment percentage, those predictors have weak positive correlation with results.

Based on information gathered thanks to correlation matrix (Figure 8) and due to trial and error, following priors were picked:

$$\alpha \sim N(\mu = 47, \sigma = 10) \quad (4)$$

$$\beta_G \sim N(\mu = 0, \sigma = 0.1) \quad (5)$$

$$\beta_{HDI} \sim N(\mu = 0, \sigma = 1) \quad (6)$$

$$\beta_U \sim N(\mu = 0, \sigma = 0.1) \quad (7)$$

$$\sigma \sim N(\mu = 1, \sigma = 0.5) \quad (8)$$

Figure 9 shows Prior distributions of priors in form of histograms.

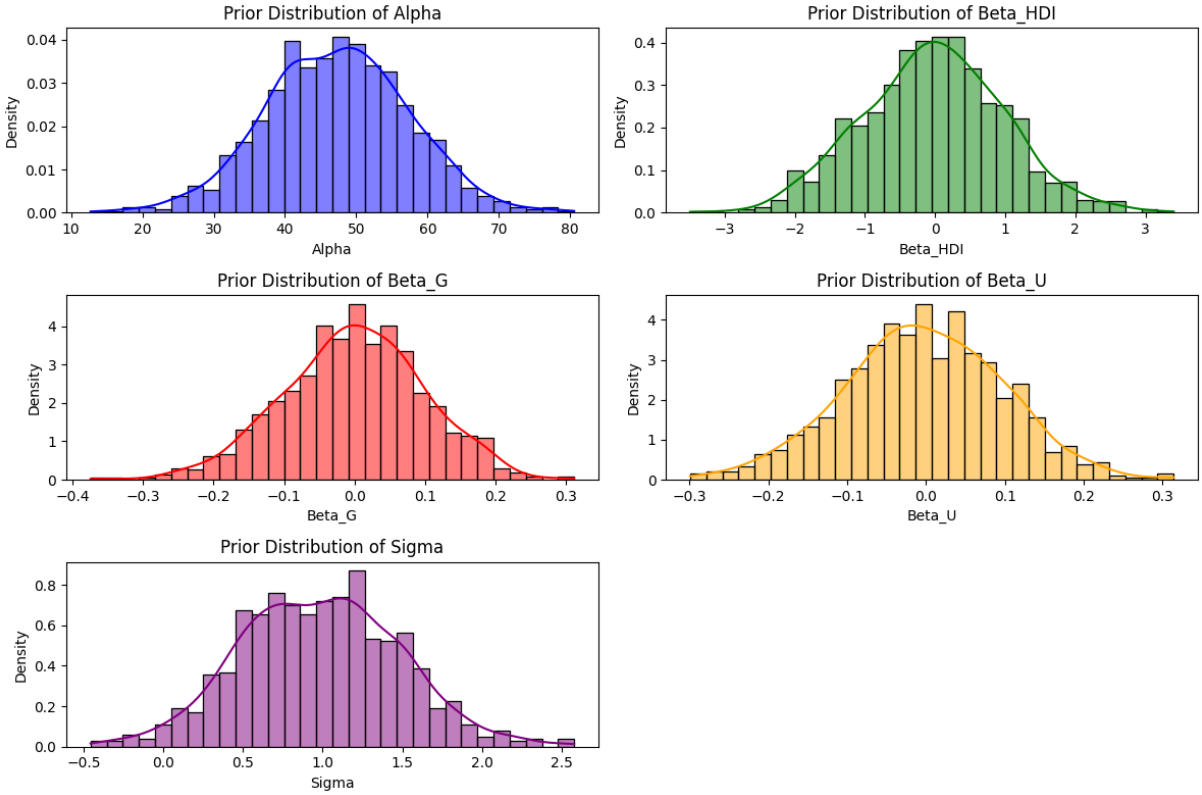

Figure 9. Histograms of Prior distributions of parameters

After putting obtained parameters into Linear regression model formula histogram was plotted. Prior predictive distribution yielded satisfactory results when compared to actual data distribution.

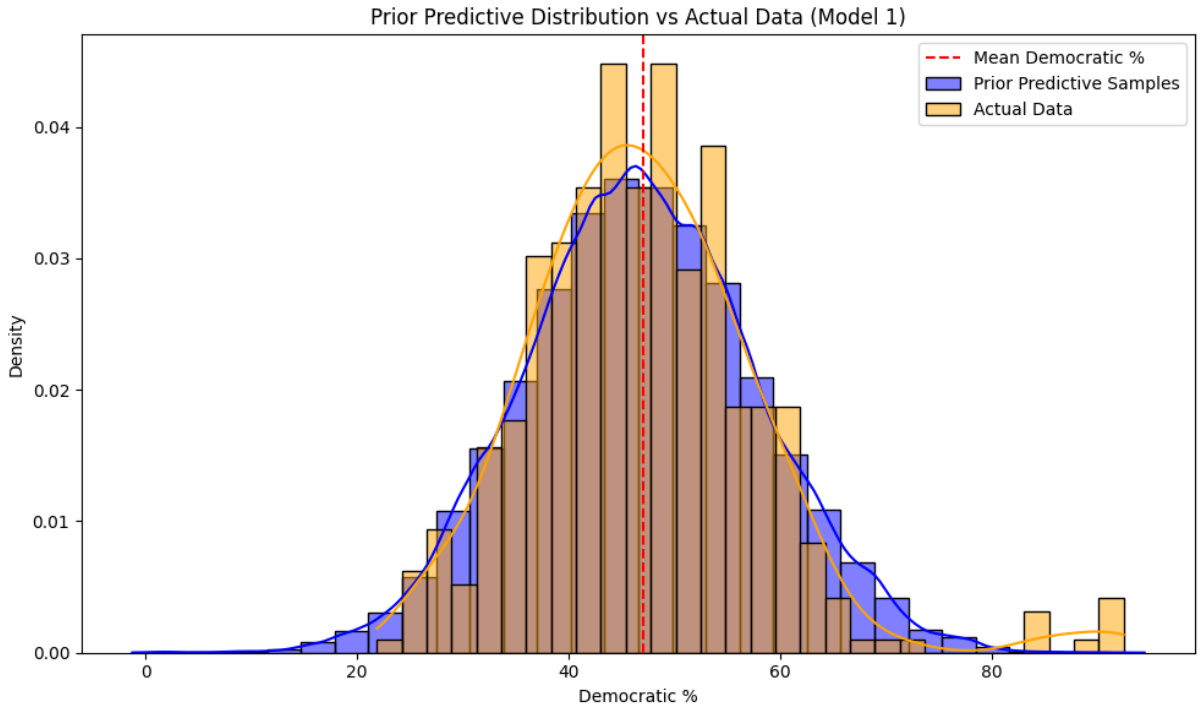

Figure 10. Comparison of Prior predictive distribution and actual data distribution

4.2 Polynomial regression model

When it came to picking priors for Polynomial regression model, this task was more complicated than picking priors for Linear regression model. However, with some assumptions it was possible to obtain a well fitted Prior predictive distribution for this model as well.

Most of overlapping parameter values (beside Unemployment percentage) were copied from linear model and for parameters next to squared predictors trial and error was applied once more. This resulted in acquiring following priors:

$$\alpha \sim N(\mu = 47, \sigma = 10) \quad (9)$$

$$\beta_{G1} \sim N(\mu = 0, \sigma = 0.01) \quad (10)$$

$$\beta_{G2} \sim N(\mu = 0, \sigma = 0.001) \quad (11)$$

$$\beta_{HDI1} \sim N(\mu = 0, \sigma = 1) \quad (12)$$

$$\beta_{HDI2} \sim N(\mu = 0, \sigma = 0.1) \quad (13)$$

$$\beta_{U1} \sim N(\mu = 0, \sigma = 0.01) \quad (14)$$

$$\beta_{U2} \sim N(\mu = 0, \sigma = 0.001) \quad (15)$$

$$\sigma \sim N(\mu = 1, \sigma = 0.5) \quad (16)$$

That is when it was decided to use predictors only up to second power, since picking priors for succeeding powers would require once again trying to find values for priors which would be a good fit. Computation power also played big role, since it took much longer to calculate third power Polynomial regression model compared to model using second power. And lastly second power Polynomial regression model provided a good enough representation of data.

Figure 11 shows Prior distributions of priors in form of histograms.

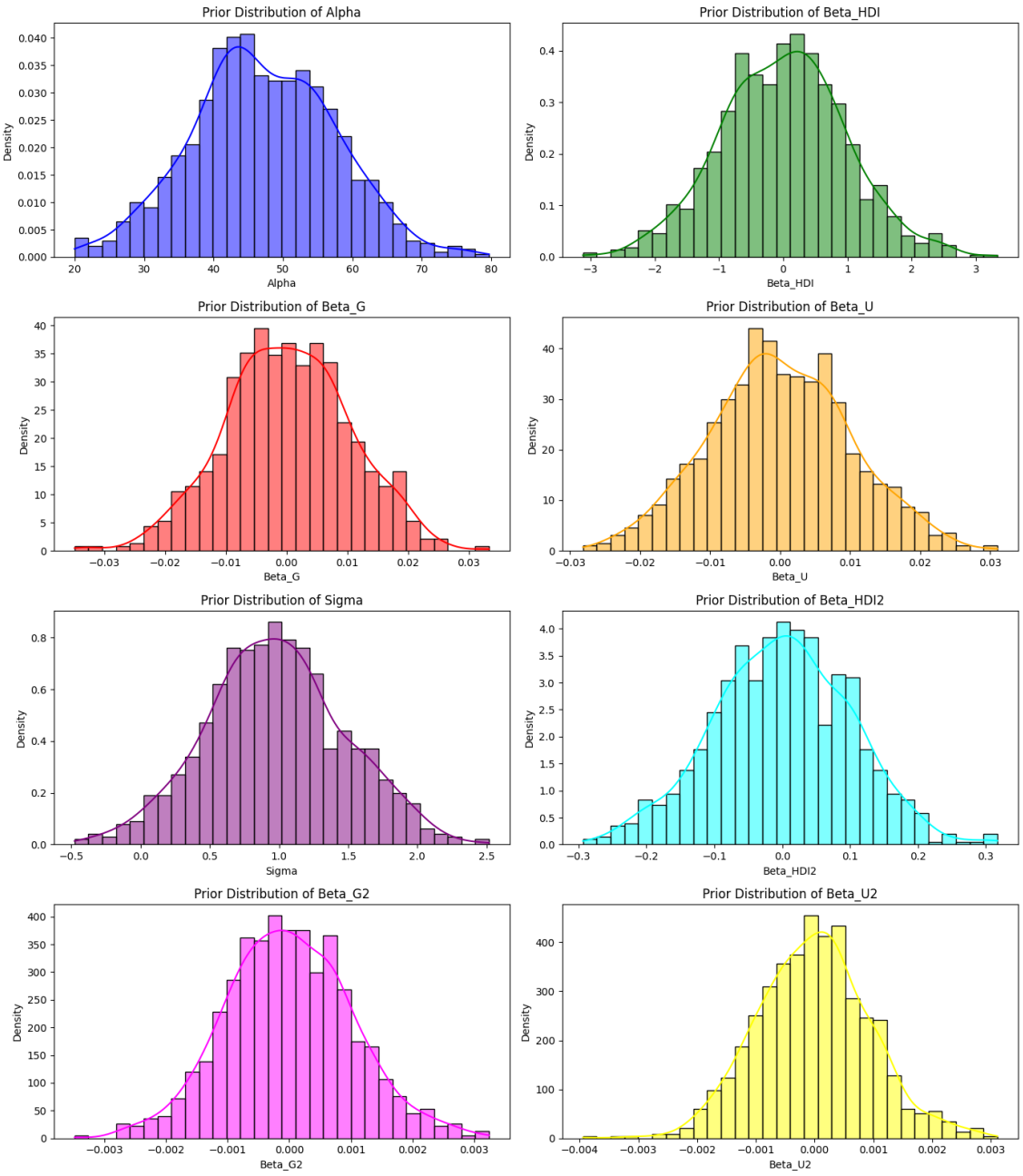

Figure 11. Histograms of Prior distributions of parameters

After putting obtained parameters into Polynomial regression model formula following histogram was plotted. Prior predictive distribution also yielded satisfactory results when compared to actual data distribution.

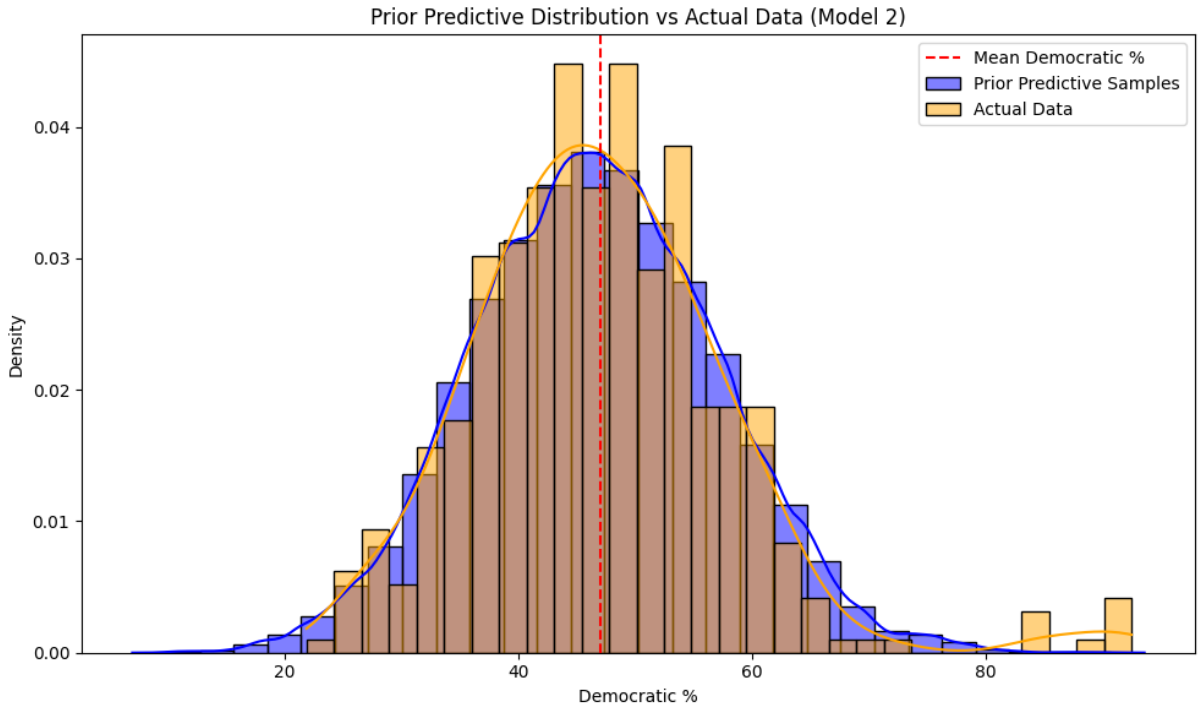

Figure 12. Comparison of Prior predictive distribution and actual data distribution

5 Posteriors

5.1 Linear regression model

Distribution of parameters was quite concentrated with following standard deviation values:

$\sigma(\alpha) = 1.3556$

$\sigma(\beta_G) = 0.0201$

$\sigma(\beta_{HDI}) = 0.9787$

$\sigma(\beta_U) = 0.0867$

$\sigma(\sigma) = 0.1722$

Figure 13 shows Posterior distributions of parameters in form of histograms.

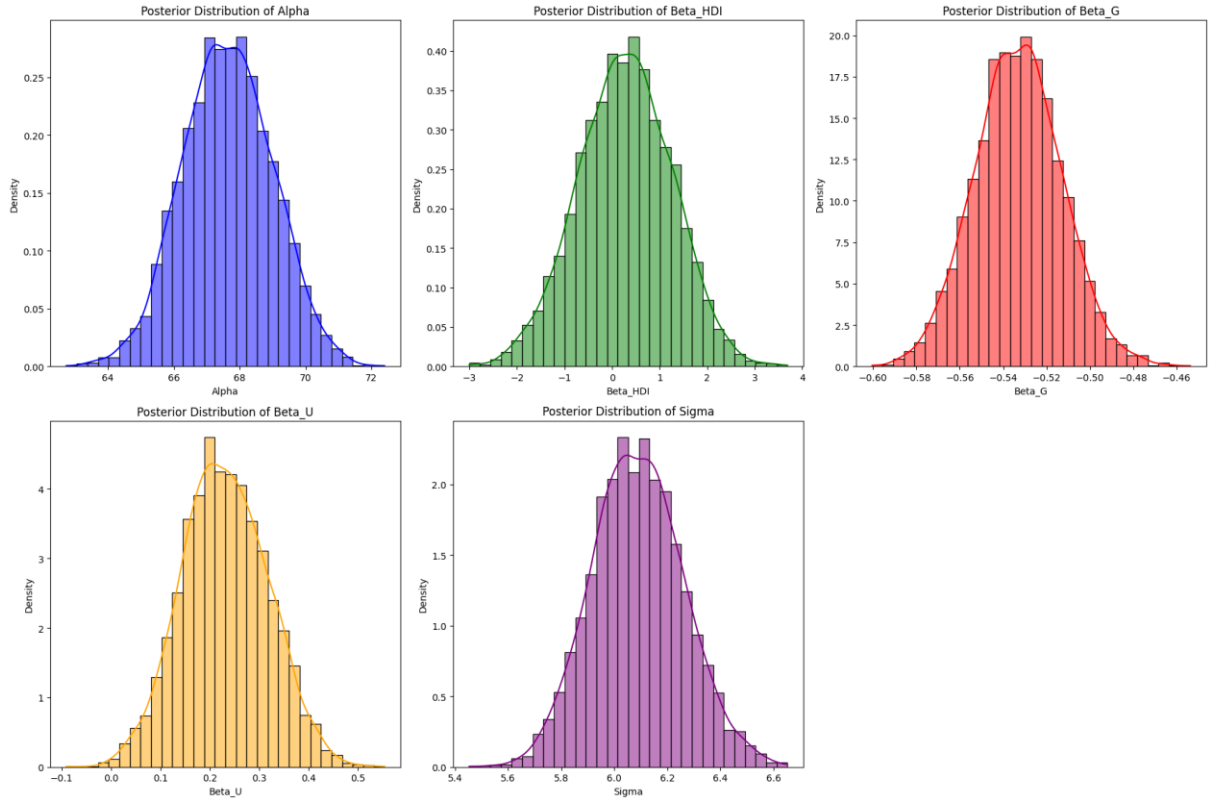

Figure 13. Histograms of Posterior distributions of parameters

Results obtained by Posterior predictive distribution yielded satisfactory results when compared to actual data distribution.

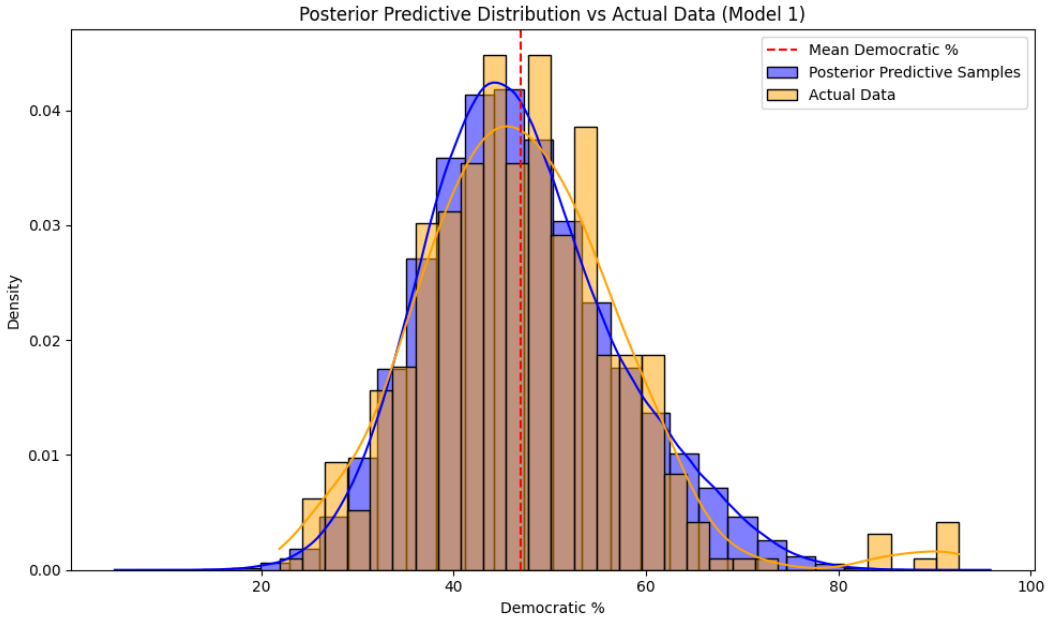

Figure 14. Comparison of Posterior predictive distribution and actual data distribution

When Linear regression model had access to all data points, data generated by the model quite well represented actual data, but this model wasn't great at representing bigger fluctuations in data.

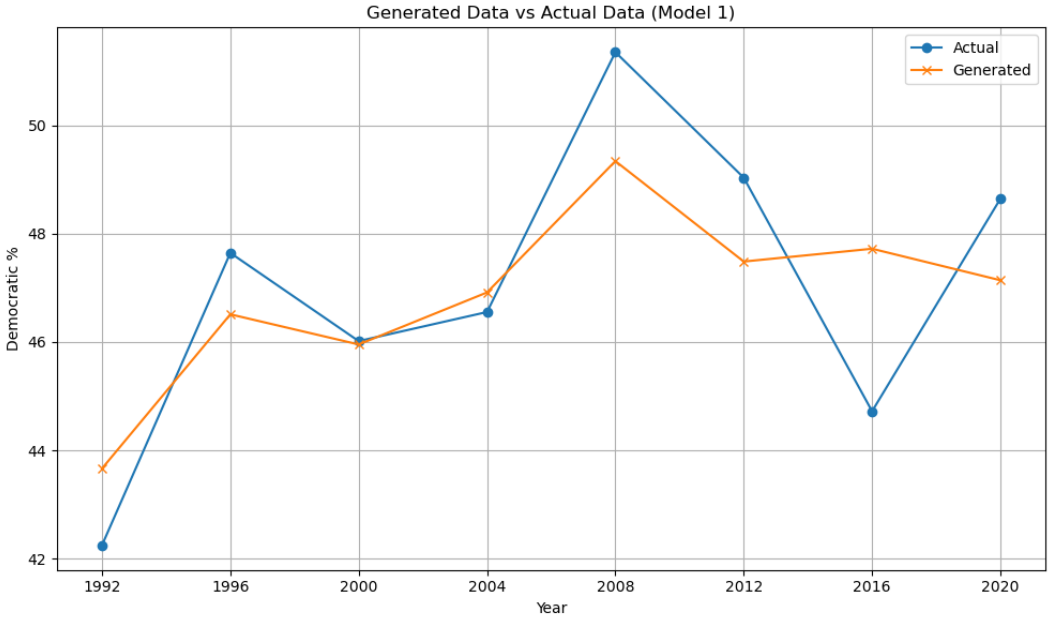

Figure 15. Comparison of generated results of presidential elections in US for each election and actual results

When it came to predicting results without specific data point, there was quite a fall off (1-2%) compared to when model had access to all data points.

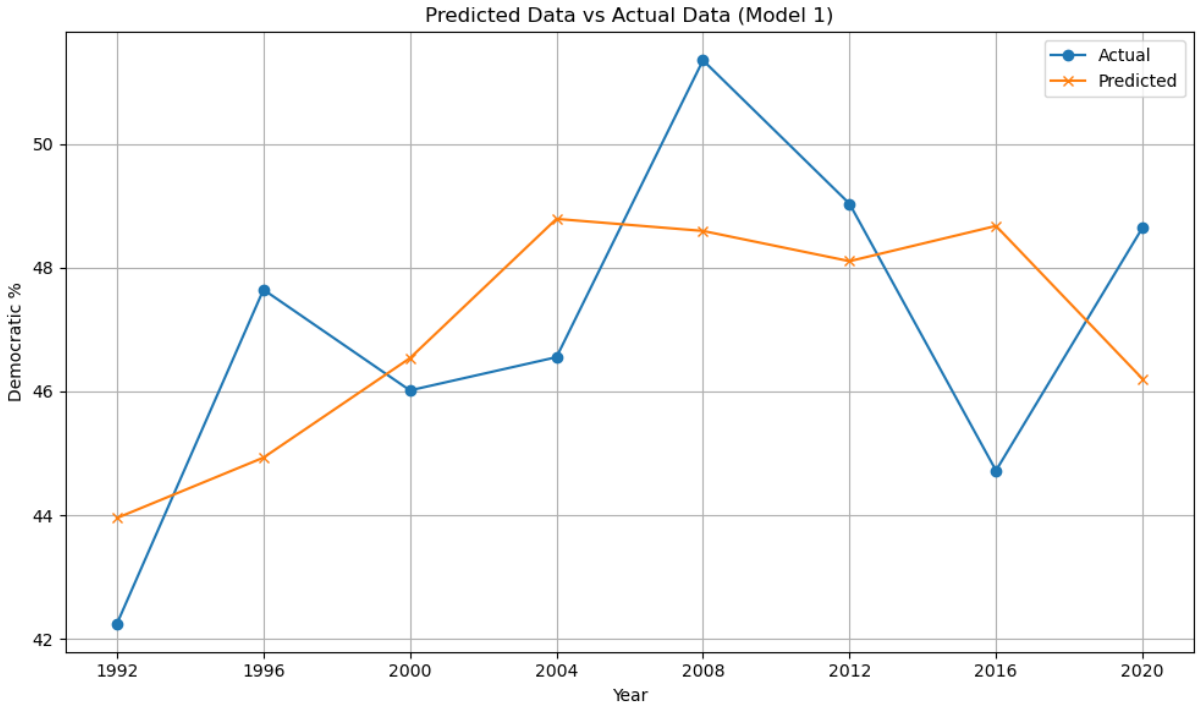

Figure 16. Comparison of predicted results of presidential elections in US for each election and actual results

5.2 Polynomial regression model

Sampling for this model took much longer than for the Linear regression model. Distribution of parameters also was quite concentrated with following standard deviation values:

$\sigma(\alpha) = 1.1093$ $\sigma(\beta_{HDI2}) = 0.1005$

$\sigma(\beta_{G1}) = 0.0101$ $\sigma(\beta_{U1}) = 0.0101$

$\sigma(\beta_{G2}) = 0.0003$ $\sigma(\beta_{U2}) = 0.0010$

$\sigma(\beta_{HDI1}) = 0.9811$ $\sigma(\sigma) = 0.1742$

An interesting fact that was noticed, standard deviation of parameters, which were multiplied by squared predictors, were much smaller than standard deviation of parameters multiplied by first power of predictors (it could have been over 33 times smaller).

Most probable suspected cause was, because small fluctuations of data cause big changes in results, since values are squared, then those parameters must be more precise.

Also compared to linear regression model most overlapping standard deviations of parameters had similar values. Only difference being $\sigma(\beta_{G1})$, which was twice as small as $\sigma(\beta_G)$.

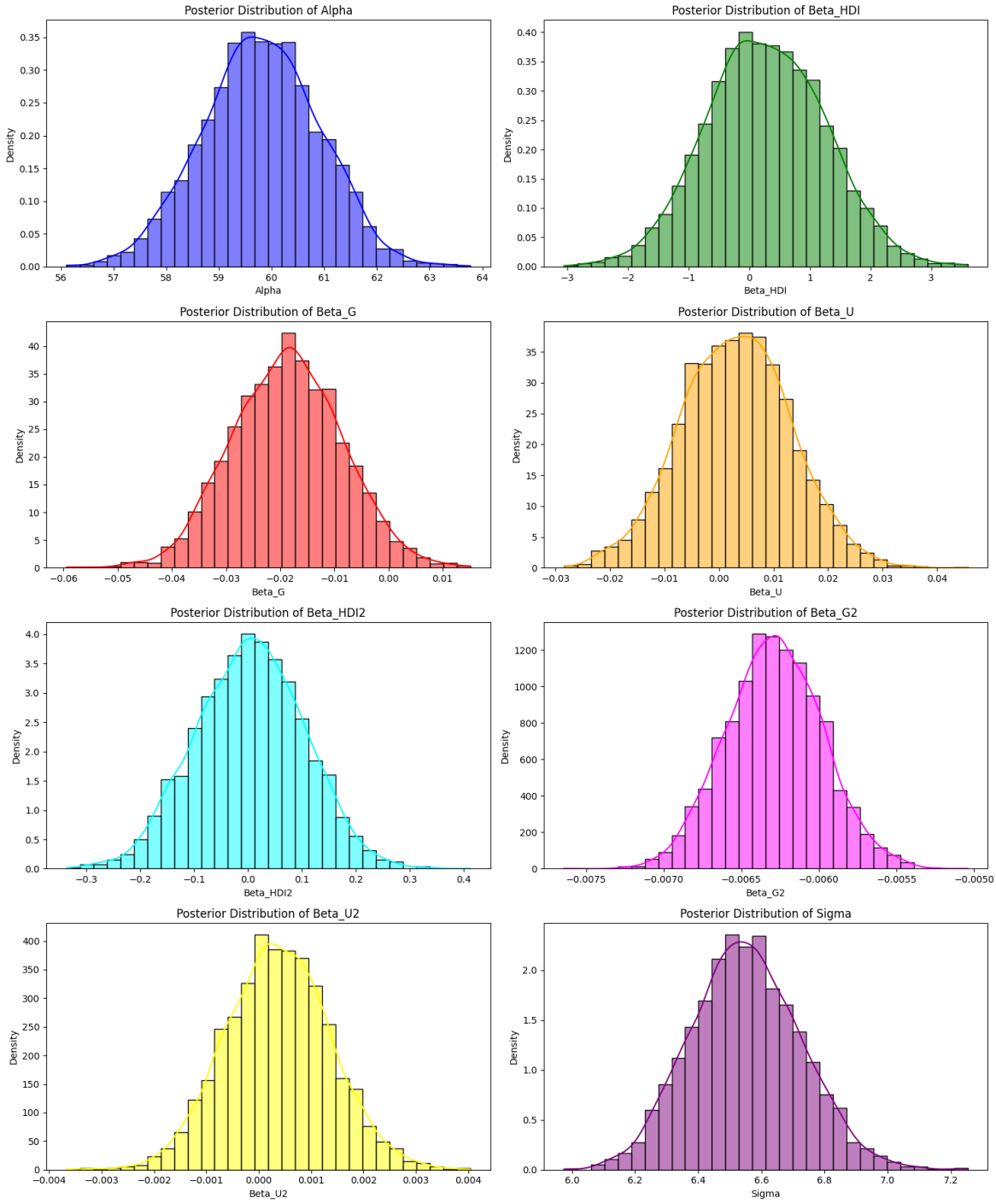

Figure 17. Histograms of Posterior distributions of parameters

Results obtained by Posterior predictive distribution yielded satisfactory results when compared to actual data distribution.

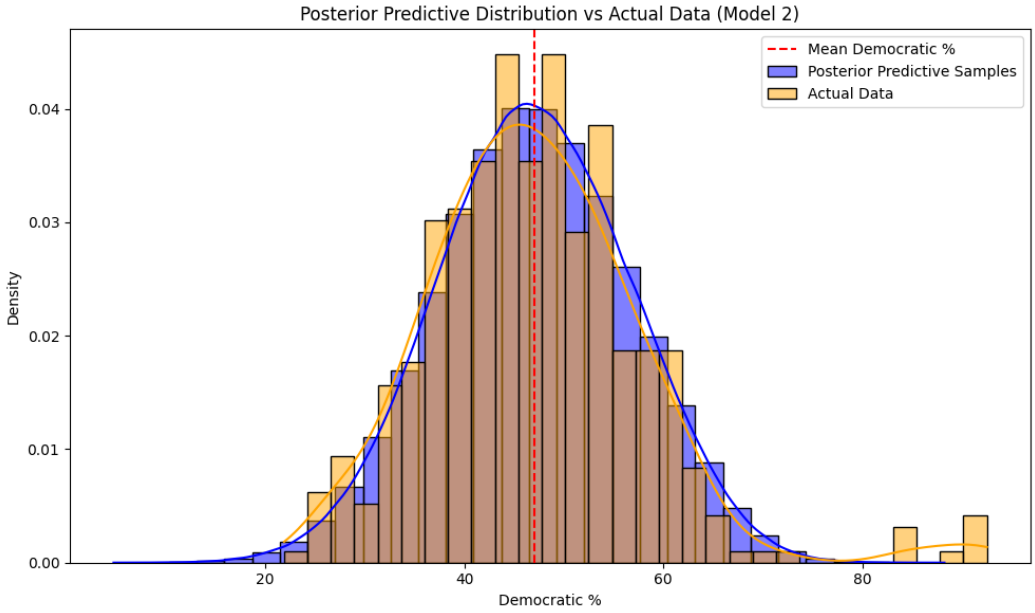

Figure 18. Comparison of Posterior predictive distribution and actual data distribution

When Polynomial regression model had access to all data points it was better at representing variation in data, but still wasn't perfect at generating all data points.

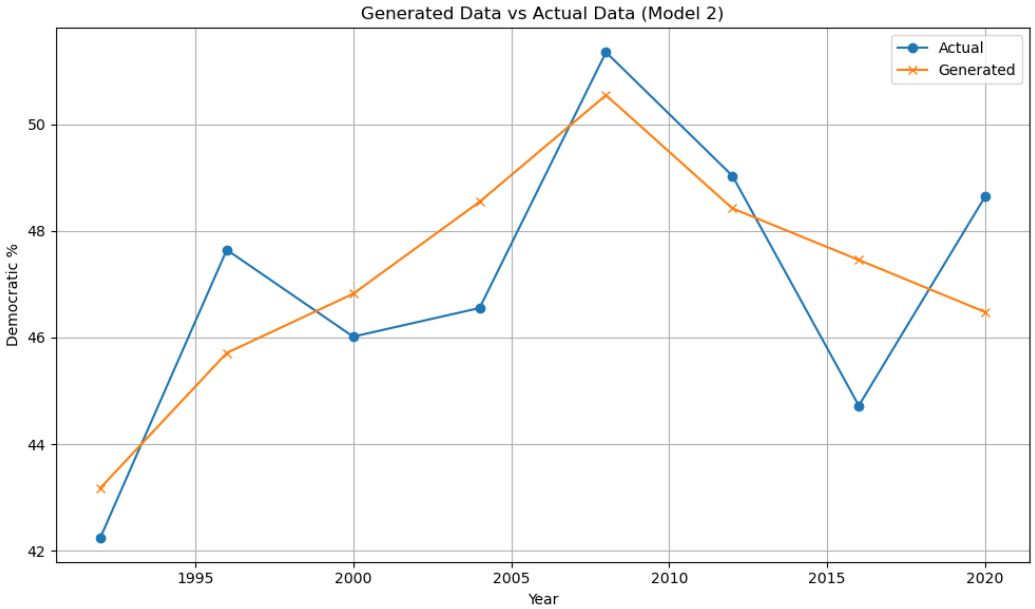

Figure 19. Comparison of predicted results of presidential elections in US for each election and actual results

When it came to predicting results without specific data point, Polynomial regression model lost its capability to accurately represent variation in data.

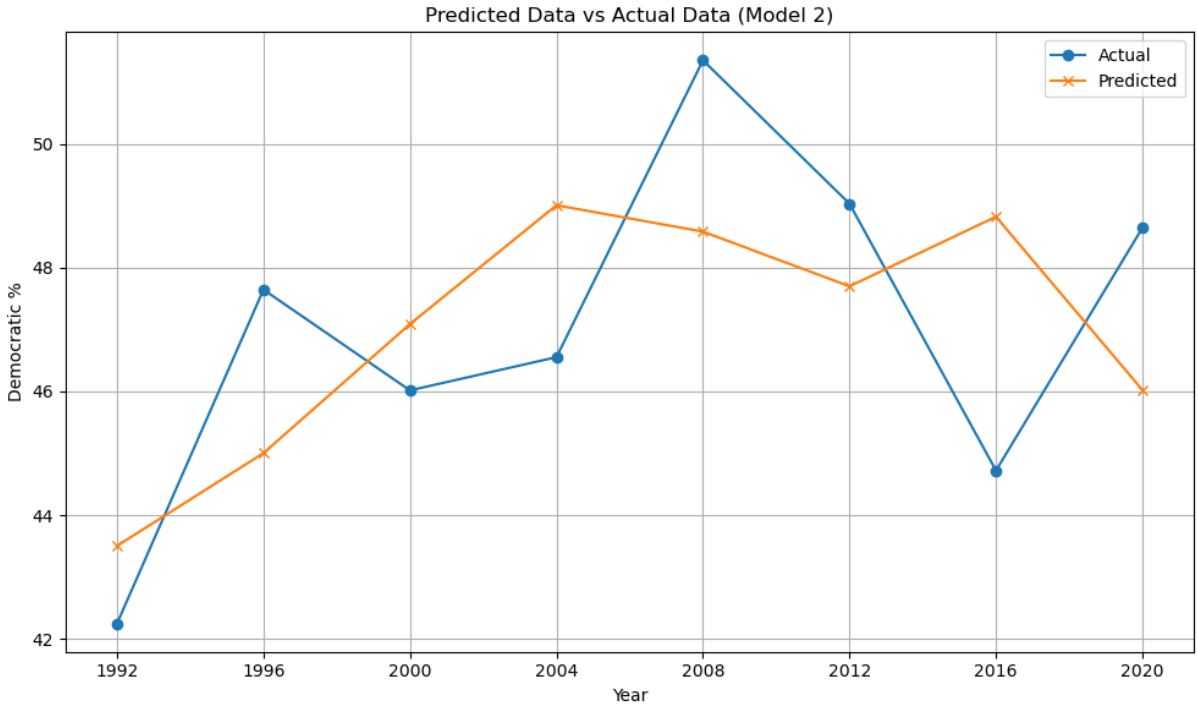

Figure 20. Comparison of predicted results of presidential elections in US for each election and actual results

6 Model comparison

Following criterions for Bayesian models were used:

- Watanabe-Akaike Information Criterion (WAIC), which works by computing the variance for each data point and then summing those variances up (one of two approaches).
- Leave-One-Out (LOO), which leaves one sample out while predicting (for models with lots of data points this change is negligible).

6.1 Comparing Linear regression model and Polynomial regression model

Both WAIC and LOO returned same results for all cases, so following statement applies to both criterions.

Even though there is quite an overlap, Linear regression model performed better than Polynomial regression model. There were no warnings.

It was hard to believe at first, that Linear model performed better than Polynomial model, but after closer inspection of graphs it became quite noticeable.

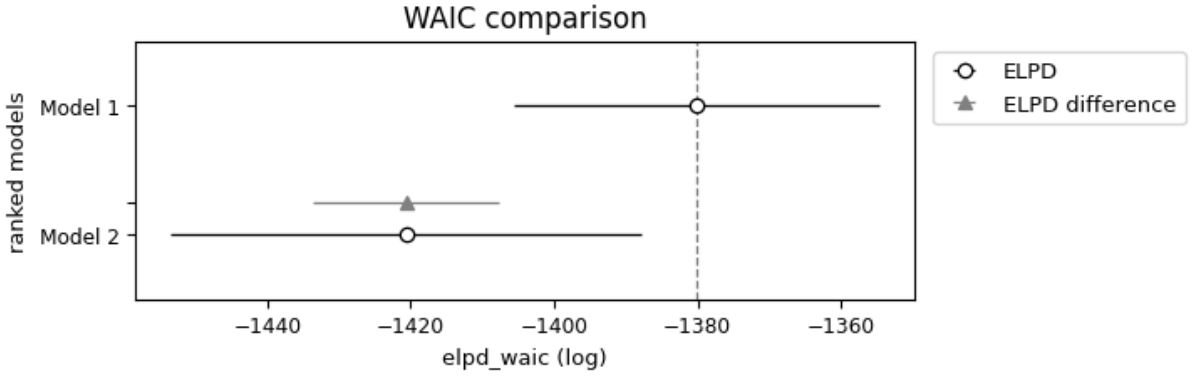

Figure 21. Comparison of Linear regression model (top) and Polynomial regression model (bottom) using WAIC

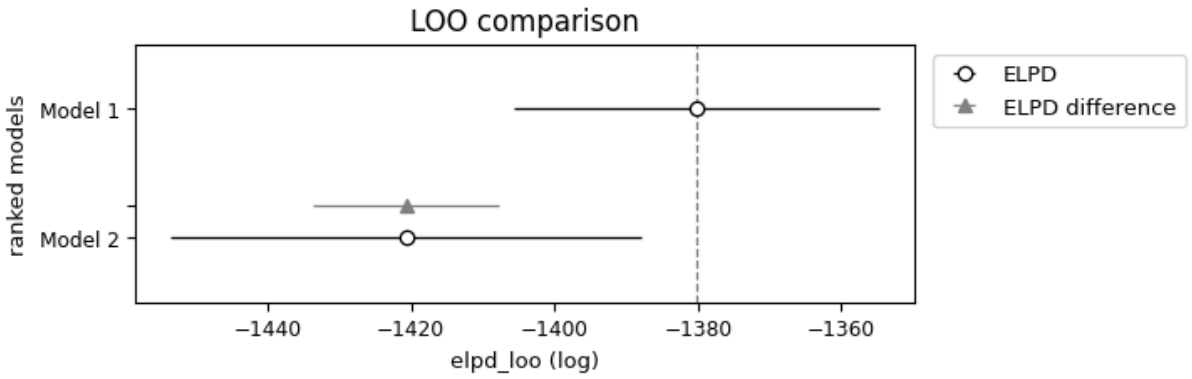

Figure 22. Comparison of Linear regression model (top) and Polynomial regression model (bottom) using LOO

6.2 Comparing models with different numbers of predictors - linear regression

It was also decided to check if using all the predictors was necessary. First model used only Human Development Index, to the second model gun ownership percentage was added and lastly unemployment percentage.

Figure 23 and 24 show using only Human Development Index was not enough, but 2 predictor model was able to perform almost as well as 3 predictor model, when it came to Linear regression.

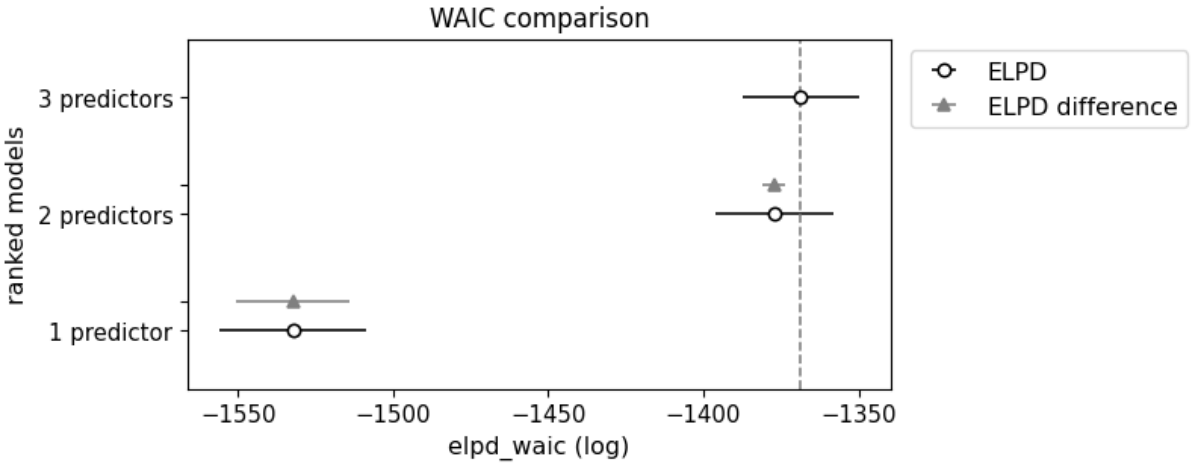

Figure 23. Comparison of Linear regression models with different number of predictors using WAIC

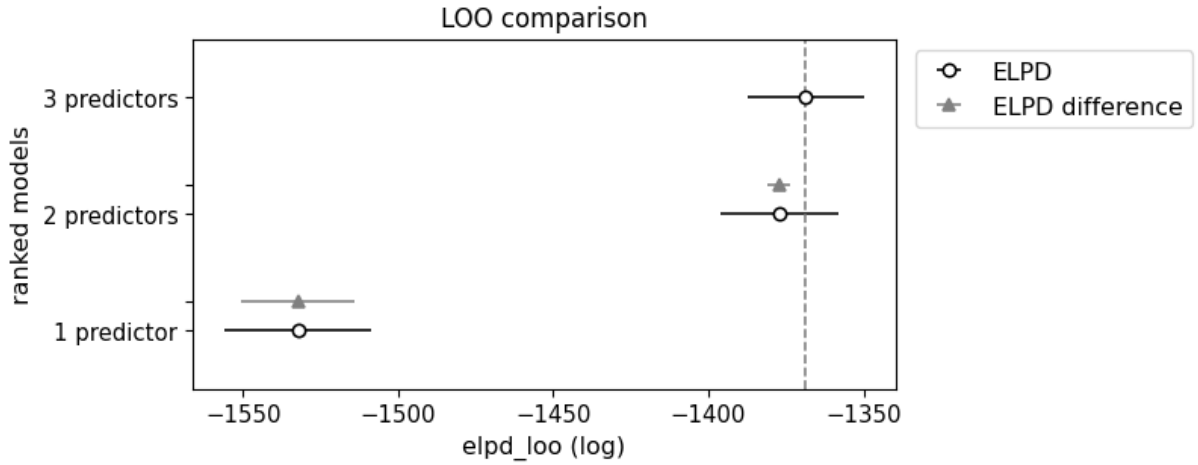

Figure 24. Comparison of Linear regression models with different number of predictors using WAIC

6.3 Comparing models with different numbers of predictors - polynomial regression

Surprisingly, when it came to Polynomial regression model, Human Development Index and gun ownership percentage were enough, and addition of unemployment percentage didn't improve model score.

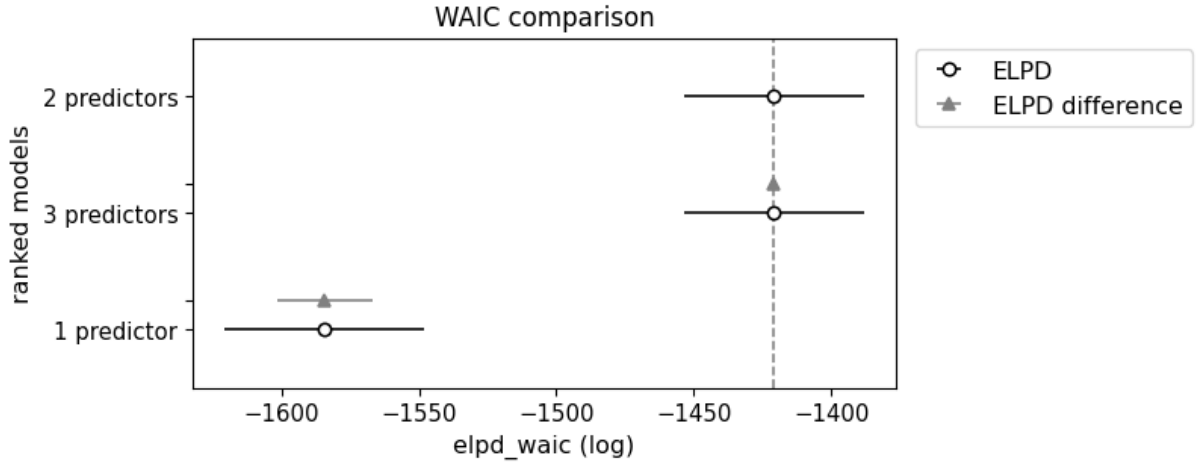

Figure 25. Comparison of Polynomial regression models with different number of predictors using WAIC

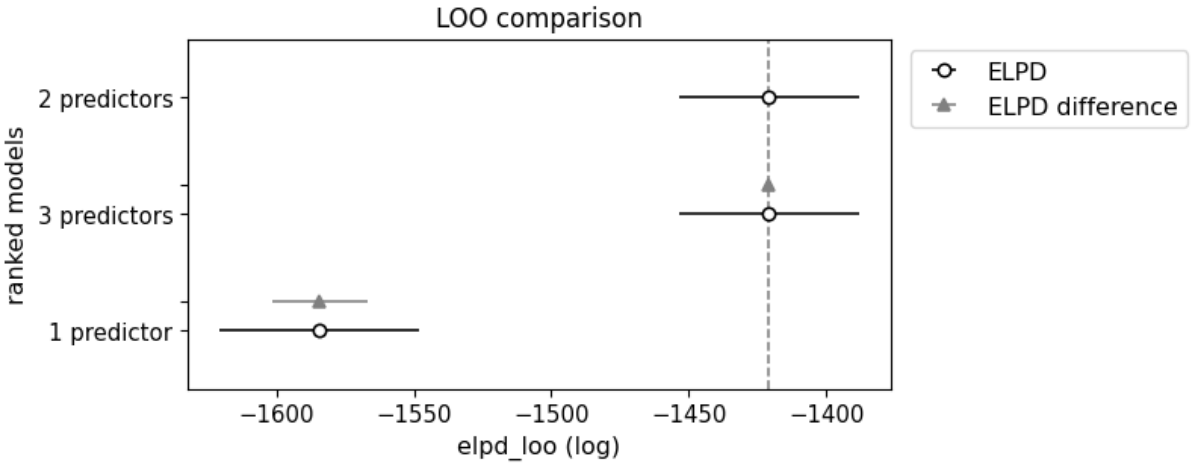

Figure 26. Comparison of Polynomial regression models with different number of predictors using LOO

7 Summary

Predicting the US presidential election was an interesting subject to work with. However, the topic itself was complicated, and it was not possible to find many articles or papers concerning it on the internet. Those that could be found were much more complicated than what timeframe allowed for.

One of the biggest surprises was that the Polynomial regression model did worse than the Linear regression model. It was assumed, that it would perform better, since there were some fluctuations in data, that was supposed to be recreated.

In the end, models did quite well - they obtained a maximum of around 5% error in terms of predictions, which could be seen as acceptable. The prediction could have been more successful, if more complex model had been used (for example hierarchical model). Use of different set of predictors also could have improved model performance.