**Predicting rates of**

**presidential action on petitions**

**for clemency and the proportion**

**of petitions granted that are**

**pardons using linear regression models**

**Loren Meyer**

**Submitted to Prof. Brinkerhoff**

**Machine Learning**

**Nov. 24, 2020**

**Abstract**

Presidents of the United States have the constitutional and statutory power to pardon prisoners as well as to commute their sentences, and grant respites and remissions. Both the ratio of petitions for clemency granted to petitions closed or denied by presidents, and the percentage of petitions granted that result in pardons, have varied considerably over the 20th and early 21st centuries. Here we investigate the utility of a linear regression model for predicting the percentage of applications for clemency that a given president in a given year will grant, and the percentage of such applications granted that will be pardons. Ultimately we find that the model offers only slight improvements in prediction over a baseline model.

**Background and data collection**

Article II, Section 2 of the Constitution gives the President the power to “grant reprieves and pardons for all offenses against the United States, except in cases of impeachment.”[[1]](#footnote-11588) The U.S. Supreme Court has since ruled that the power extends to the power to “remit fines and forfeitures” and “commute sentences” and that the power is “unlimited” and “not subject to legislative control.”[[2]](#footnote-9433) Clemency is exercised for a variety of reasons, including on “legal or technical” grounds, as an exercise of compassion or mercy, due to a judgment that a prisoner has reformed, or due to electoral considerations.[[3]](#footnote-2114) A survey of Presidential exercises of clemency between 1900 and 1993 found that presidents during this period granted on average more than 200 petitions per year.[[4]](#footnote-5332)

“Positive actions” that can be taken by the President with respect to clemency petitions include pardons, in which a crime is entirely forgiven with full restoration of civil rights; commutations, in which a lesser sentence is substituted for the original one; remissions, in which a financial or other penalty is reduced; and respites, in which imposition of a sentence is delayed, often pending a review of the fairness of the trial and sentencing.[[5]](#footnote-13704)

From Kaggle (a well-known repository of datasets), I obtained two data files which collectively contained information about presidential actions on clemency from fiscal year 1900 through fiscal year 2017, indexed by year and by President, based on information provided by the Office of the Pardon Attorney.[[6]](#footnote-3384)

**Data cleaning and identifying target variables**

The dataset included some complementary patterns of missing information: for some years, it only included a total of petitions denied and petitions closed without presidential action; for some years, it did not include this total, but included separate totals for petitions denied and petitions closed without presidential action. These fields were combined so as to retain the maximal information possible that could be found for every year in the dataset.

The dataset did not include totals for petitions received for every year. Since preliminary investigation of supplementary data sources (e.g., the data presented in Ruckman’s *Presidential Studies Quarterly*) paper revealed minor inconsistencies with the Kaggle dataset, no attempt was made to supplement the missing information with information drawn from other sources; instead, the analysis ignored the information on petitions received.

The data for 1900-1967 included totals for pardons, commutations, respites, and remissions. The data for 1967-2017 included totals for pardons, commutations, and remissions, but not respites. In both cases, the sum of these actions was identified as a total number of positive actions taken. Since this was known to vary as a function of petitions received[[7]](#footnote-881), and information on petitions received was incomplete, and since presidential inaction also varied greatly from one administration to another[[8]](#footnote-24678), the percentage of petitions that received positive action among those that received positive action or were denied or closed without presidential action and the proportion of petitions that received positive action that were specifically pardons were identified as target variables less likely to vary as a function of unavailable information.

One year/president combination included very little data. A quick web search[[9]](#footnote-26966) indicated that the president did not in fact hold office at any time during the indicated fiscal year, and so that particular data point was excluded from the analysis.[[10]](#footnote-459)

For several years, no positive action, whether pardons or some other form, was taken on clemency. Since years were treated as a separate independent variable, to avoid creating gaps the missing values for the dependent variable pardons as a proportion of total positive action were substituted using mean imputation.[[11]](#footnote-25093)

**Data exploration and identification and extraction of predictor variables**

A chart of the primary target variable (percentage of resolved clemency cases resulting in positive action) over time suggested the possibility of splitting the dataset into an early and a late set. A date (between 1967 and 1968) for splitting the dataset for the prediction of this variable was suggested based on maximizing the total reduction in standard deviation for this variable in each subset, scaled by the proportion of points assigned to that subset.[[12]](#footnote-19233) A scatter plot of this variable by political party of the president which included a division of the data into early and late points suggested that the effect of political party varied considerably between the early and late datasets, and so the data points were ultimately divided into 4 bins: Early Republicans, Early Democrats, Late Republicans, and Late Democrats. This categorization was carried over to the efforts to predict the secondary target variable (proportion of positive actions resulting in pardons), although in this case between 1938 and 1939 was identified as the most informative splitting point based on the criterion of reduction in total proportional standard deviation.

No attempt was made to condition on the identity of the president other than based on the comparatively obvious criterion of political party, which had already been identified[[13]](#footnote-23887) as a likely explanatory variable. Although Ruckman[[14]](#footnote-14359) noted success by previous researchers in using a subjective identification of presidential governing styles to predict proportions of petitions for clemency granted, no attempt was made to use this identification since it was ad hoc and subjective and since the original categorization did not include Presidents Clinton, George W. Bush, or Obama. Such categories as Christian denomination or ethnicity were rejected due to lack of a compelling causal hypothesis and the likelihood of mistaking random variation of the (high-variance) target variables that is coincidentally correlated with these predictors for a reproducible effect.[[15]](#footnote-31405)

A box plot[[16]](#footnote-31791) suggested that the fiscal year *modulo* 4 (that is, the situation relative to the presidential election cycle) had potential to be informative with regard to the primary target variable; this same information was carried over for prediction of the secondary target variable.

Finally, since there was little else information available, and since the curve visually suggested a slight long-term decline in both the proportion of cases resolved that were positive and in the proportion of positive actions that were pardons, the years themselves were used as an independent variable in addition to the other variables indicated. These were expressed as a fraction of the total number of years since 1899.[[17]](#footnote-28466)

**Model selection and fitting**

A design matrix was created after encoding the party/date combination and election cycle identifier as one-hot variables.[[18]](#footnote-16132) Since it could be inferred that a president in a given year was a Late Republican if and only if he was not an Early Republican or a Democrat (Early or Late), and since it could be inferred that the fiscal year *modulo* 4 was 3 if and only if it was not 0,1, or 2, these categories were left out as is standard practice to avoid the ”dummy variable trap.” Since this was done, an intercept could be fitted to the data without creating the redundancy involved if each category were to be included.[[19]](#footnote-27277)

Since every feature was based on the year, there was a potential for high correlation between features. “The intercorrelation between explanatory variables is termed as ‘multicollinearity.’” This is a problem that can result in the failure of standard numerical analysis methods to find a best-fit line as well as misleading results that amplify the effects of small variations (such as measurement errors) in the target data.[[20]](#footnote-7920) Two rules of thumb suggesting unacceptably high multicollinearity are a correlation coefficient between explanatory variables greater than 0.70, or a design matrix condition number[[21]](#footnote-10110) greater than 30. A design matrix with a condition number less than 15 is thought to be ”well-conditioned” and unlikely to result in drastically different results to the equation Az=b if the values of b are slightly changed.[[22]](#footnote-25319) Singular values less than 1 are also indicative of a matrix likely to amplify both small variations in the target vector and numerical errors due to floating-point arithmetical imprecision.[[23]](#footnote-25651)

Since correlation is a measure of linear dependence, there was unsurprisingly little correlation between the fiscal year (scaled as indicated) and the situation relative to the 4-year election cycle. There were high correlations (correlation coefficients greater than 0.60) between the fiscal year and the categories based on the time period and presidential party, although presumably lower than if the party were not a component of the categorization. However, these did not exceed the 0.70 rule of thumb value given by Kim. The design matrices (one for each target variable since one used the post-1967 date split point and the other used the post-1938 date split point) both had condition numbers less than 15, indicating well-conditioned matrices, and one singular value close to 1 but none less than 1.

The proposed model, chosen for simplicity and transparency, is summarized by the equation ŷi=b+Σall j Ai,j Ѳj, where ŷi is the predicted value of the target variable for observation *i*, b is a constant intercept term, Ai,j is the observed value of independent variable *j* for observation *i*, and Ѳj is a calculated coefficient specific to that independent variable.

**Model fitting and evaluation**

The model was given a standard linear algebra formulation in terms of an *i x n+1* design matrix A where entry A i,j is the observed value of independent variable *j* of the *n* (in this case, n=7) independent variables where *j<=n* and 1 where *j=n+1*. A vector *x* of coefficients augmented by an intercept term was then fitted to the equation Ax=y to find a fit that minimized the root mean square error[[24]](#footnote-19309) between the calculated value Ax= ŷand the observed value y. To calculate this vector, we used Python’s NumPy library’s built-in least-squares calculator.[[25]](#footnote-16547)

The data were randomly split into a training set used to calculate the model coefficients and a test set used to test how robust the calculation was. To evaluate the success of the model, we compared its performance on the test set to a baseline model based on the equation ŷi=b (that is, one that minimized RMSE by predicting the mean regardless of the features of the observed datapoint. We found a modest reduction of approximately 17% compared to the baseline for predicting the primary target variable, and a reduction of approximate 5% for predicting the secondary target variable (note, though, that as noted above the improvements are perhaps underestimated since the imputation method used had the effect of particularly improving the performance of this baseline model).

Ridge regression reduces variance at the expense of adding bias, thus preventing a model from being fitted to noise in the training data. Moreover, “Ridge regression provides a means of addressing the problem of [multi-]collinearity without removing variables from the original set of independent variables.”[[26]](#footnote-4977) ... although the rules of thumb noted above generally suggest multicollinearity is not a big problem in this case, the relatively high correlations of certain independent variables and the single low singular value for the design matrix suggest caution. Ridge regression finds the best fit to the equation Ax=y subject to the proviso that the values of the components of x should not be particularly high. More precisely, it minimizes a specified linear combination of the MSE and the Euclidean norm of x. A ridge regression model was fitted based on the observation that minimizing the (penalized) loss function is equivalent to augmenting the original design matrix with a conformable identity matrix multiplied by a scalar and fitting it to an observation vector supplemented by an appropriate number of zeros.[[27]](#footnote-10428) Exploration of plausible penalties for large values of components of x suggested that ridge regression in this case did not improve accuracy with respect to the training set. This, the unusual superior performance on the (smaller and less varied) test data than on the training data, and the evident (in the charts of predicted vs. actual values over time) performance that is particularly poor with regard to outliers in the observations, suggest[[28]](#footnote-19134) that, far from a too complex model that is fitted well to the training data but does not generalize well, we have an ”underfit” model that is not complex enough to capture all the variance.

**Conclusions**

With caution, linear regression coefficients can be interpreted as indicating the relative import of different explanatory factors.[[29]](#footnote-11314) Though the correlation of independent variables means more careful statistical methods would need to be brought to bear to quantify the importance of these factors, inspection suggests a high likelihood of a long-term decline in the ratio of convicts granted clemency to those whose petitions were denied or closed without executive action, and a likelihood that Democrats were more likely to grant clemency than Republicans.[[30]](#footnote-27380) It also appears likely that time period has a bigger effect on both dependent variables than party.

Although it (inevitably) provided a modest increase in explanatory capability over the crude baseline, the attempt to use a simple linear regression model with the identified independent variables to predict the percentage of cases resolved resulting in clemency in a given fiscal year for a given president, and to predict the percentage of positive clemency actions that were pardons, must be accounted a failure. The range of variation of the errors was considerably greater than many of the values for the target variable.

The appearance that the model is underfit to the data suggests that finding an improved model requires adding complexity. This could mean adding new independent variables or replacing existing ones. If the current ones are used, they appear to be suited perhaps to a decision tree or Naïve Bayes approach. Visual inspection suggests that the performance of a linear regression model might look considerably better if outliers are excluded, so another approach could be using an approach like Naïve Bayes (or even logistic regression) to identify likely outliers on each end, and adding another Boolean term to the design matrix that adds a term based on the residuals of the original linear regression model in cases so identified. Of course, ultimately neural networks, though they have the drawback of being computationally expensive, time-consuming, and relatively opaque, are capable in principle of solving pretty much any machine learning problem, given sufficient data.[[31]](#footnote-15817) And, of course, if as in this case there is little data to go on, it will inevitably be difficult to make a data-driven prediction.

1. Ruckman, P.S. “Executive Clemency in the United States.” *Presidential Studies Quarterly* 27, no. 2 (Spring 1997): 251. Accessed online. [↑](#footnote-ref-11588)
2. Ruckman. “Executive Clemency.” Page 253. Accessed online. [↑](#footnote-ref-9433)
3. Ruckman. “Executive Clemency.” Pages 256-258. [↑](#footnote-ref-2114)
4. Ruckman. “Executive Clemency.” Page 261. [↑](#footnote-ref-5332)
5. Portman, Janet. “Presidential Clemency: Pardons, Commutations, and Reprieves.” Accessed online 11/24/20 at [Presidential Clemency: Pardons, Commutations, and Reprieves | CriminalDefenseLawyer.com](https://www.criminaldefenselawyer.com/resources/presidential-clemency-pardons-commutations-and-reprie). [↑](#footnote-ref-13704)
6. [Presidential Pardons, 1900-2017 | Kaggle](https://www.kaggle.com/doj/presidential-pardons) [↑](#footnote-ref-3384)
7. Ruckman. “Executive Clemency.” Page 258. [↑](#footnote-ref-881)
8. Ruckman. “Executive Clemency.” Pages 259-260. [↑](#footnote-ref-24678)
9. [Richard Nixon - Wikipedia](https://en.wikipedia.org/wiki/Richard_Nixon) [↑](#footnote-ref-26966)
10. When data is missing for non-random reasons, it is generally better to use imputation or statistical analysis that explicitly accounts for the missing data rather than to remove data points from the analysis. However, when there are indications that collection of a single data point may have been a mistake, it can sometimes be justifiable to exclude that point. Rob Smith, University of Montana Computer Science Department, personal communication. [↑](#footnote-ref-459)
11. Imputation is generally preferred over deletion of data points when data is not missing at random. Imputation of the mean has the advantage of being easy and of not making assumptions about the validity of a particular model as in other methods such as maximum likelihood imputation. Mean imputation of target variables, however, tends to result in an underestimation of errors, particularly in the lowest-variance models including linear regression models. Rob Smith, UM Computer Science, personal communication. Note that in this case the underestimation of error will be greatest in the case of the baseline model which returns precisely zero error in this case, and so the comparative advantage of the multivariable linear regression model will if anything be underestimated. [↑](#footnote-ref-25093)
12. This is a principled method for partitioning datasets based on the value of a numerical variable. Javier Perez-Alvaro, University of Montana Mathematics Department, personal communication. [↑](#footnote-ref-19233)
13. e.g., in Ruckman’s *Presidential Studies Quarterly* article, page 261. [↑](#footnote-ref-23887)
14. “Executive Clemency.“ Pages 258-259. [↑](#footnote-ref-14359)
15. This is a common danger when introducing new independent variables, especially in the absence of reasons to expect a cause-and-effect relationship. Note that the introduction of new independent variables can only result in reduction of errors between the model predictions and the training data, and that this does not mean that the model will perform better on a new data point. Douglas Brinkerhoff and Rob Smith, University of Montana Computer Science Department, personal communication. [↑](#footnote-ref-31405)
16. This is a good method of visualizing at once the average value and the range of variation of a given variable, and is especially useful when comparing the value of that variable conditioned on a categorical independent variable. Brian Steele, University of Montana mathematics department, personal communication. [↑](#footnote-ref-31791)
17. Standardization of this sort can make more sense than *z*-standardization or scaling by standard deviation when the only other variables are categorical. Brian Steele, UM Math, personal communication. [↑](#footnote-ref-28466)
18. This is the most appropriate way to encode information that is categorical and not continuous or ordinal. Brian Steele, UM Math, and Douglas Brinkerhoff, UM Computer Science, personal communication. [↑](#footnote-ref-16132)
19. ”If several systems of classes are involved the best procedure is to delete one dummy variable from each system.” Suits, Daniel. ”Use of Dummy Variables in Regression Equations.” *Journal of the American Statistical Association* 52, no. 280 (1957). Page 548. See also the general discussion of this problem, pages 548-551. [↑](#footnote-ref-27277)
20. Kim, Jong Hae. ”Multicollinearity and Misleading Statistical Results.” *Korean Journal of Anesthesiology* 72, no. 6 (2019): pages 558-569. Quote appears on page 558. [↑](#footnote-ref-7920)
21. ”The condition number of a matrix A is a numerical measure of the accuracy attainable in the solution of the linear system Az=b.” Casella, George. ”Condition Numbers and Minimax Ridge Regression Estimators.” *Journal of the American Statistical Association* 80, no. 391 (1985). Page 754. [↑](#footnote-ref-10110)
22. Kim. ”Multicollinearity and misleading statistical results.” Pages 560-561. [↑](#footnote-ref-25319)
23. Javier Perez-Alvaro, UM Math Department, personal communication. [↑](#footnote-ref-25651)
24. The root mean squared error (RMSE) is a monotonically increasing function of the mean squared error (MSE), meaning that any approach that minimizes the MSE also minimizes the RMSE, but is particularly appropriate to linear regression because it directly measures the standard deviation of the errors. Brian Steele, UM Math Dept., personal communication. [↑](#footnote-ref-19309)
25. This calculator which is based on the singular value decomposition is more numerically stable than the ordinary least-squares calculation based on the normal equations, and not dependent on correct hyperparameter selection or the assumption of approximate convexity as with gradient descent. Javier Perez-Alvaro, UM Math Dept., personal communication. [↑](#footnote-ref-16547)
26. McDonald, Gary. ”Ridge Regression.” *WIREs Computational Statistics* 1, no. 1 (no date, accessed online 2020). Pages 93-100, quote on page 93. [↑](#footnote-ref-4977)
27. Javier Perez-Alvaro, UM Math Department, personal communication. [↑](#footnote-ref-10428)
28. Javier Perez-Alvaro and Brian Steele, UM Math Dept., and Douglas Brinkerhoff, UM Computer Science Dept., personal communication. [↑](#footnote-ref-19134)
29. Brian Steele, UM Mathematics Department, personal communication. Caution must be taken to ensure that (as in this case) the features of the data used to compute these coefficients are comparably scaled, and to take into account the interaction between independent variables that are correlated. [↑](#footnote-ref-11314)
30. Both of these conclusions are consistent with the direct calculations of total positive actions on clemency conditioned on party and date made by Ruckman with reference to the 1900-1993 period; the results with respect to party can be found on page 261 of his *Presidential Studies Quarterly* article while the long-term decline over time can be found on page 264. [↑](#footnote-ref-27380)
31. Personal observation. [↑](#footnote-ref-15817)