THE 2022 ELECTION IN THE UNITED STATES: RELIABILITY OF A LINEAR REGRESSION MODEL

Jan Kalina – Petra Vidnerová – Miroslava Večeř

Abstract

In this paper, the 2022 United States election to the House of Representatives is analyzed by means of a linear regression model. After the election process is explained, the popular vote is modeled as a response of 8 predictors (demographic characteristics) on the state-wide level. The main focus is paid to verifying the reliability of two obtained regression models, namely the full model with all predictors and the most relevant submodel found by hypothesis testing (with 4 relevant predictors). Individual topics related to assessing reliability that are used in this study include confidence intervals for predictions, multicollinearity, and also outlier detection. While the predictions in the submodel that includes only relevant predictors are very similar to those in the full model, it turns out that the submodel has better reliability properties compared to the full model, especially in terms of narrower confidence intervals for the values of the popular vote.

Key words: elections results, electoral demography, linear regression, reliability, variability

JEL Code: C21, C18, D72

Introduction

The last election to the United States House of Representatives took place on November 8, 2022. Traditionally, the elections in the USA are thoroughly analyzed and discussed as direct messages of voters to their politicians and the popular vote is interesting to be estimated for clusters of voters according to their demographic characteristics (Lytle et al., 2018).

For practical implications of a linear regression model, which predicts the popular vote based on demographic characteristics, it is natural to ask about the statistical reliability of the obtained model. The question about reliability (consistency or precision of measurements if repeated by the same or similar conditions) of a trained statistical model is general and should be answered any time when analyzing (not only) demographic data by means of statistical methods. The statistical meaning of the concept of reliability is connected to small variability of predictions; the small variability (and resistance against small perturbations of data) may be achieved at the price of biased estimation (Breneman et al., 2022). In the context of linear regression modeling, the least squares estimator is equipped with a number of available specific diagnostic tools and verifying reliability includes to study confidence intervals, outlier detection, or even interpretation of the results (Zigerell, 2022).

Still, verifying the reliability of a statistical model seems to be an insufficiently discussed topic even in monographs on statistical methods (Arkes, 2023). Previous demographic studies understood reliability as replicability of the findings using highly similar data (Matanda et al., 2014) or goodness of fit of the data with the assumed probabilistic model. Other applications that paid intensive attention to reliability include the study of demographic models for a honeybee colony of Mandal and Maity (2022), who stressed the importance of sensitivity (robustness) analysis for understanding the effects of the parameters. The work focused on stochastic models (branching processes) for predicting the dynamics in time. The sensitivity was thoroughly inspected also in the study of Rabitti and Borgonovo (2020), who investigated the effect of demographic factors in annuity models in the field of life insurance.

The unique feature of this paper is the focus on assessing reliability of the obtained model. Section 1 describes the election to the U.S. House of Representatives. Section 2 describes the analyzed data, i.e. the results of the 2022 election to the U.S. House of Representatives. The analysis predicts (explains) the popular vote in connection with some important demographic characteristics on the state-wide level. First, a linear regression model is fitted (Section 3) and a relevant subset of variables is found (Section 4). Then, the aim is to decide which of the two models is more suitable in terms of reliability. As the reliability criteria, we consider multicollinearity (Section 5), outlier detection (Section 6), and cconfidence intervals for predictions (Section 7).

1 The U.S. election process

The United States Congress consists of two mutually equal chambers and represents the highest legislative body in the United States of America, which are the House of Representatives, as the lower house, and the Senate, as the upper house. The Seat of Congress is in Washington, D.C. After years of English colonial rule and the Declaration of Independence on July 4, 1776, the United States Constitution was signed on September 17, 1787. The first U.S. Constitution shows the strength of each member state in Article I, Section I. The Constitution of the United States placed all legislative power in the hands of the Congress. The House of Representatives is established according to Article I, Section II of the U.S. Consitution of representatives who are elected by the people of the several states every second year (National Constitutional Center, 2023). A citizen can be elected if he meets the following conditions: age minimum of 25 years, representative is a U.S. citizen for a minimum of 7 years, and is an inhabitant of the state in which the election happens (Jonáš, 2008). The same requirements are also maintained in the current version of the United States Constitution. The first Congress of the United States met for its first session in New York on March 4, 1789. Four years later, the number of state representatives in Congress increased to a total of 325, based on the results of the census. Currently, for election to the House of Representatives, the United States is diversified into 435 congressional districts, with a population of approximately 760 000 and for each of those districts, the representative is elected for a period of two years to the House of Representatives. In addition to these districts, there are territories that send only their delegate to the House of Representatives, who is not gifted with the right to vote.

2 Data description

We consider the results of the United States election to the House of Representatives in the year 2022. In this section, the response as well as 8 considered predictors $X_1, \dots, X_8$ corresponding to selected (mainly demographic) characteristics of the individual 50 states of the USA are described, while District of Columbia is not considered in this study. The response Y corresponds to the percentage of popular votes for Republican candidates in the election to the House of Representatives in November 2022. For each of the variables, Table 1 gives the publicly available source of the data. All the variables are continuous. The percentages are taken as values between 0 and 100. We use R software for all the computations.

- X_1 = percentage of African Americans in the state population in 2015.
- X_2 = percentage of Hispanic and Latino population in the state population in 2012.
- X_3 = population density as the number of inhabitants per square kilometer in 2015.
- X_4 = median age in years in 2020.
- X_5 = percentage of individuals with a bachelor's or higher degree in the state population in 2021.
- X_6 = divorce rate for people at the age of 30 (the year is not specified) obtained as the percentage of divorced marriages among all marriages.
- X_7 = weekly church attendance defined as the percentage in the state population of those who attend a church, synagogue or mosque once a week or almost every week, as estimated in 2014.

- X_8 = percentage of individuals adherent to Protestant Christianity in the state population in 2014.

Tab. 1: Sources of data (8 predictors and the response)

Variable	Source of the data
X_1	https://en.wikipedia.org/wiki/Demographics_of_the_United_States
X_2	https://en.wikipedia.org/wiki/Demographics_of_Hispanic_and_Latino_Americans
X_3	https://en.wikipedia.org/wiki/List_of_states_and_territories_of_the_United_States_by_population_density
X_4	https://en.wikipedia.org/wiki/List_of_U.S._states_and_territories_by_median_age
X_5	https://en.wikipedia.org/wiki/List_of_U.S._states_and_territories_by_educational_attainment
X_6	https://www.zippia.com/research/divorce-by-30-by-state/
X_7	https://en.wikipedia.org/wiki/Church_attendance
X_8	https://en.wikipedia.org/wiki/List_of_U.S._states_and_territories_by_religiosity
Y	https://en.wikipedia.org/wiki/2022_United_States_House_of_Representatives_elections

3 Fitting the linear regression

We consider the standard linear regression model in the form

$$Y_i = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_p X_{ip} + e_i, \qquad i = 1, \ldots, n, \tag{1}$$

with $p=8$ and use the least squares estimates of the parameters. The coefficient of determination equals $R^2 = 0.78$. Table 2 presents R^2 values for (1) and also for other models described below. The values of MSE (mean square error) reported in Table 2 were evaluated within a 5-fold cross-validation. Graphical visualizations of the data (not shown here) reveal the linear model to be meaningful and justifiable.

Before we proceed to a discussion of reliability of the model (1), let us mention only two aspects in the interpretation of the effect of individual predictors (regressors) on the response. Our first remark is related to education. Strongly Democratic-leaning states with large values of X_5 suffer from overeducation. On the other hand, strongly Republican-leaning states are rural states with a lower average income with a complicated access to financing university studies and also with pressures on young people to establish their families soon. Our second remarks concerns the population density X_3. Although the plot of Y against X_3 clearly shows that Y is typically decreasing with an increasing X_3, the estimate of β_3 in (1) is positive. Such controversial result is a consequence of multicollinearity in (1), which leads e.g. to overestimated predictions of Y for the urban states of New England.

4 Model building

Statistical regression modeling is typically accompanied by model choice (model building), which attempts to find a submodel that contains only the relevant predictors. By applying backward selection based on standard t-tests on the model (1), we arrived to the submodel with 4 predictors in the form

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_5 X_{i5} + \beta_7 X_{i7} + e_i, \quad i = 1, \dots, n. \tag{2}$$

In other words, the two most relevant predictors are education and church attendance and two other relevant ones (but with a weaker association) the percentages of African Americans and Hispanic/Latino population. The same model (2) with $R^2 = 0.75$ was found by an alternative approach to dimensionality reduction, which considered all the possible submodels and found the best one as that with the smallest value of Akaike information criterion (AIC). We do not perform other dimensionality reduction procedures; the popular principal component analysis (PCA) is very unsuitable here because the predictors are non-commensurate. All the following reliability issues are considered for the full model (1) as well as for the submodel (2).

5 Multicollinearity

Because of large correlations in the set of predictors, it is clear that multicollinearity represents an issue in (1). The condition number of the matrix $X^T X$ is commonly used as a measure of the sensitivity of the least squares estimate to perturbations of the input data and multicollinearity is understood as a serious issue if the conditional number exceeds 30. The condition number of the matrix of predictors is defined as the ratio of the largest eigenvalue of $X^T X$ to the smallest eigenvalue. It is equal here to a very high value 4986.7 for $p=8$. In (2), the condition number drops to 57.1; this is not a safe value, but is much improved compared to the value in the full model (1). Some of the predictors in the submodel have also high correlations with other predictors, e.g. the percentage of ethnic minorities in the population with church attendance.

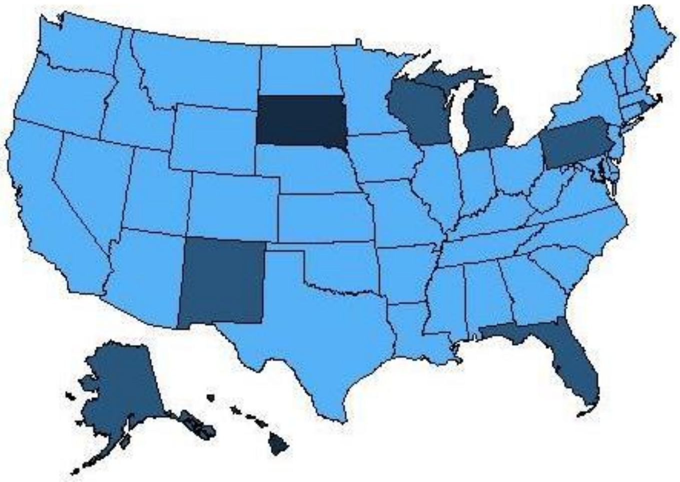
6 Outlier detection

In the full model (1), there are two severe outliers: South Dakota, where the Democratic candidate withdrew before the election in 2022, and Hawaii with a very specific demographic structure. Table 2 gives summary results also after omitting the 2 severe outliers from the data, i.e. for *n=48.* In addition, the model (1) yields a wrong prediction of the strongest party in 6 states. These are (in alphabetical order) Alaska, Michigan, New Mexico, Florida, Pennsylvania, and Wisconsin and are shown in Figure 1. The study of outliers in the submodel yields analogous results. The leverage scores, which are formally defined as diagonal elements of the projection matrix $X(X^T X)^{-1}X^T$, allow finding the states that are outliers in terms of the predictors (when not taking Y into account).

7 Confidence intervals

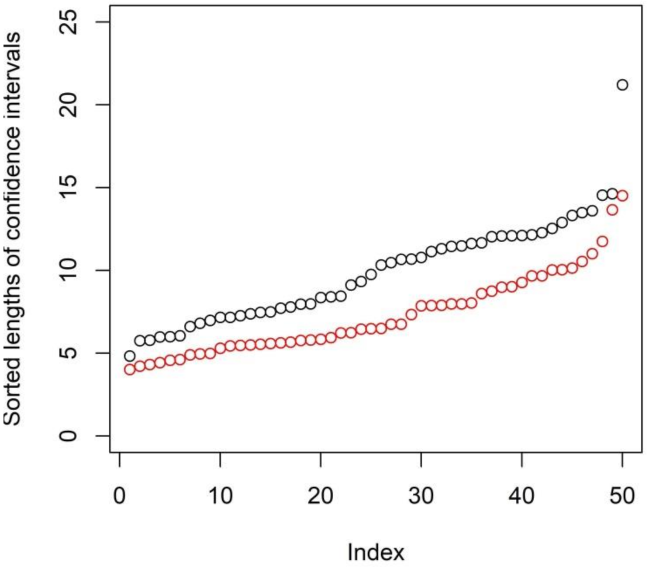
Confidence intervals for the response (for an individual state) are evaluated under the assumption of normally distributed random errors and under homoscedasticity. Sorted lengths of the confidence intervals for both (1) and (2) are shown in Figure 2. The very same patterns are obtained also for the sorted leverage scores of Section 5; to explain this, the length of the confidence intervals is proportional to the leverage scores.

Fig. 1: Outlying states in the full model, where 2 outliers are dark (South Dakota, Hawaii) and 6 outliers with a wrong prediction of the election winner are medium dark



Source: https://en.wikipedia.org/wiki/2022_United_States_House_of_Representatives_elections

Fig. 2: Sorted lengths of confidence intervals for the full model (1) (black) and the submodel (2) (red)



Source: https://en.wikipedia.org/wiki/2022_United_States_House_of_Representatives_elections

Although it is theoretically known that multicollinearity does not affect point estimates of regression parameters, it turns out that removing multicollinearity is beneficial from the practical point of view. This is because removing multicollinearity means narrowing the confidence intervals and the width of the confidence intervals is a crucial criterion of reliability. Here, Utah turns out to be the outlying state with the widest confidence interval and Ohio and Iowa have the narrowest ones, i.e. their demographic characteristics are not outlying.

The confidence intervals could theoretically be affected by possible heteroskedasticity in the model. We emply tests of the null hypothesis of homoskedasticity, which are known as important diagnostic tools for the regression modeling (Khaled et al., 2019). The Breusch-Pagan test of heteroskedasticity yields the *p*-value *p=0.766* here so that we may conclude that heteroscedasticity does not represent an issue in the model. As a consequence,, we consider the confidence intervals to be reliable in the given situation.

Conclusion

The analysis of the election data shows a quite strong association of the popular vote with the predictors. The effect of demographic predictors on the popular vote has been known (Lytle et al., 2018), but out main contribution is the study of reliability, i.e. aspects related to the variability of predictions. We consider the submodel (2) with 4 main predictors to be more reliable than the full model (1), which is clear e.g. from Figure 2 with length of the confidence intervals. Naturally, our analysis is simplified and a comparison with the previous election results would be very useful.

Reliability issues, which are typically evaluated by means of confidence intervals of predictions, represent one of important aspects of regression modeling. While **dimensionality reduction** and **model choice** are not usually associated with reliability, they turn out to play a key role in finding a reliable model (Kalina and Rensová, 2015). Other aspects of regression modeling (multicollinearity, outlier detection) are investigated here for the election data as well; it should kept in mind here that there are several outliers not well explained by the model. Obtaining reliable results requires a careful training and tuning of the considered model and even if we consider the model to be reliable, the results are obtained with a quite high uncertainty and thus deserve to be interpreted with care. While the submodel (and not the full model) would be selected as the final model by an experienced statistician for a variety of reasons, this paper recalls that reliability is one of key reasons for selecting the submodel. The reliability aspects are studied here without using tools of robust statistics. While the approaches of nonparametric statistics have appealing properties in terms of robustness to measurement errors (Saleh et al., 2012), replacing the least squares estimator by a robust alternative would require to perform reliability verifications (Kalina, 2015).

Reliability considerations are important for every regression method, i.e. not only for the linear model used here, but also for machine learning regression tools, which are primarily focused on predictions (rather than explainability). For example, Zhang et al. (2021) performed a detailed verification of reliability for a hybrid multi-stage classification system for credit risk tasks. Let us also note that reliability should not be mistaken for accuracy, where the latter concept is usually perceived as unbiasedness (lack of bias) in the context of the analysis of measurement errors. In general, machine learning applications in demography may exploit confidence intervals (which can be always obtained by bootstrap), outlier detection, or model choice (sparsity on the level of ignoring redundant features). However, using a black-box machine learning regression method may yield results with a lower reliability compared to linear models as e.g. in Weng et al. (2019).

Acknowledgment

The work was supported by the project 21-19311S of the Czech Science Foundation.

References

Arkes, J. (2023). *Regression analysis. A practical introduction.* 2nd edn. Routledge, London.

Breneman, J.E., Sahay, C., & Lewis, E.E. (2022). *Introduction to reliability engineering.* 3rd edn. Wiley, Hoboken.

Jonáš, K., ed. (2008). *Americké právo 1886: Sbírka zákonů a výkladů právních, pro osadníky česko-americké zvláště důležitých.* College of Applied Law, Prague. (In Czech.)

Kalina, J. (2015). Three contributions to robust regression diagnostics. *Journal of Applied Mathematics, Statistics and Informatics, 11*(2), 69–78.

Kalina, J. & Rensová, D. (2015). How to reduce dimensionality of data: Robustness point of view. *Serbian Journal of Management, 10*(1), 131–140.

Khaled, W., Lin, J., Han, Z., Zhao, Y., & Hao, H. (2019). Test for heteroscedasticity in partially linear regression models. *Journal of Systems Science and Complexity, 32*, 1194–1210.

Lytle, A., Macdonald, J., Dyar, C., & Levy, S.R. (2018). Ageism and sexism in the 2016 United States presidential election. *Analyses of Social Issues and Public Policy, 18*, 81–104.

Mandal, P.S. & Maity, S. (2022). Impact of demographic variability on the disease dynamics for honeybee model. *Chaos, 32*, 083120.

Matanda, D.J., Mittelmark, M.B., Urke, H.B., & Amugsi, D.A. (2014). Reliability of demographic and socioeconomic variables in predicting early initiation of breastfeeding: A replication analysis using the Kenya demographic and health survey data. *BMJ Open, 4*, e005194.

National Constitutional Center (2023). *The United States Constitution.* [https://constitution-center.org/the-constitution]

Rabitti, G. & Borgonovo, E. (2020). Is mortality or interest rate the most important risk in annuity models? A comparison of sensitivity analysis methods. *Insurance: Mathematics and Economics, 95*, 45–58.

Saleh, A.K.M.E., Picek, J., & Kalina, J. (2012). R-estimation of the parameters of a multiple regression model with measurement errors. *Metrika, 75,* 311–328.

Weng, S.F., Vaz, L., Qureshi, N., & Kai, J. (2019). Prediction of premature all-cause mortality: A prospective general population cohort study comparing machine-learning and standard epidemiological approaches. *PLoS One, 14,* e0214365.

Zhang, W., Yang, D., & Zhang, S. (2021). A new hybrid ensemble model with voting-based outlier detection and balanced sampling for credit scoring. *Expert Systems with Applications, 174*, 114744.

Zigerell, L.J. (2023). Introducing political science students to data visualization strategies. *Journal of Political Science Education, 19*, 1–15.

Contact

Jan Kalina
The Czech Academy of Sciences, Institute of Computer Science
Pod Vodárenskou věží 2, 182 00, Prague 8, Czech Republic
& Charles University, Faculty of Mathematics and Physics
Sokolovská 83, 186 75 Prague 8, Czech Republic
kalina@cs.cas.cz

Petra Vidnerová
The Czech Academy of Sciences, Institute of Computer Science
Pod Vodárenskou věží 2, 182 00, Prague 8, Czech Republic
petra@cs.cas.cz

Miroslava Večeř
Charles University, Faculty of Law
Department of Financial Law and Financial Science
nám. Curieových 901/7, 116 40 Prague 1, Czech Republic
vecer@prf.cuni.cz