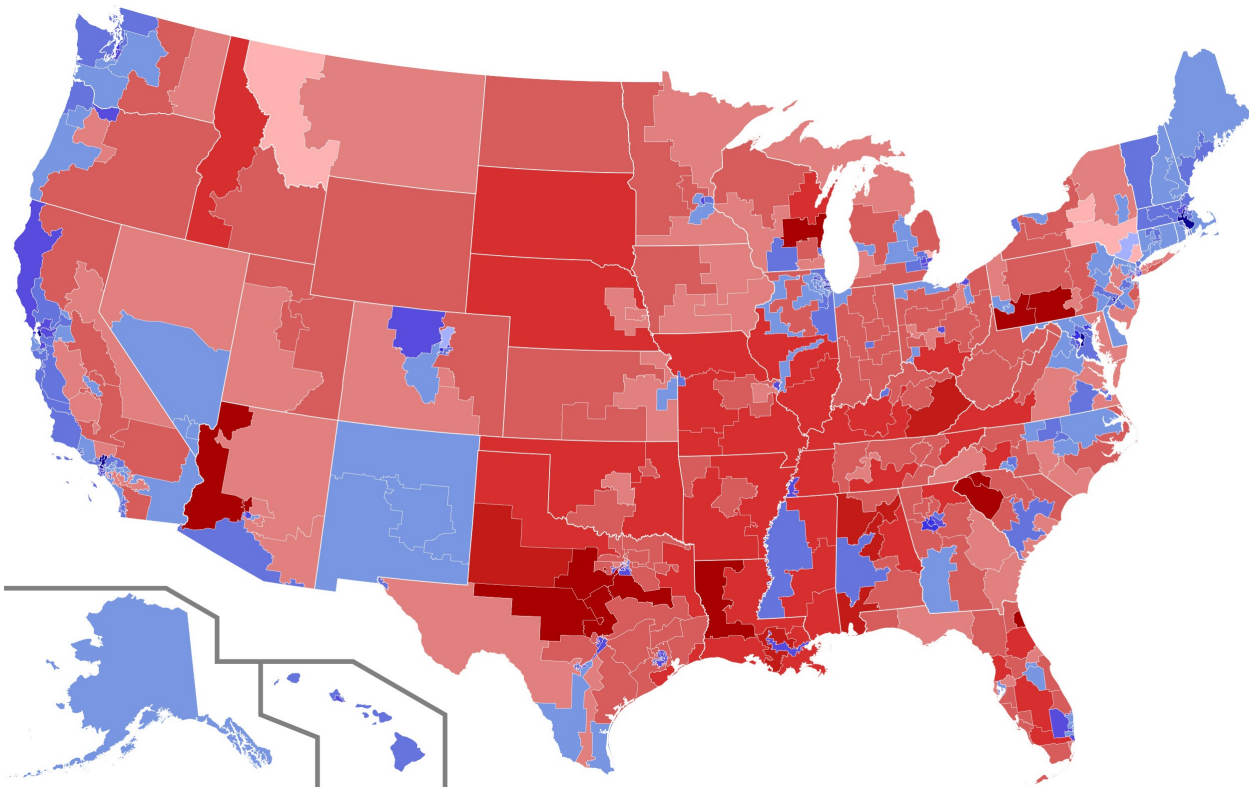
The 2022 Election in the United States

How to Verify Reliability of Linear Regression

Jan Kalina & Petra Vidnerová & Miroslava Večeř

The Czech Academy of Sciences, Institute of Computer Science, Prague

The 2022 election in the United States: results



- Election to the House of Representatives on November 8, 2022
- Red: Republican party (obtained 222 seats)
- Blue: Democratic party (obtained 212 seats)
- Source: wikipedia

The data for individual 50 states

- Y = percentage of popular vote for the Republican Party
- X_1 = percentage of African American population 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 obtained as the percentage of divorced marriages among all marriages.
- X_7 = weekly church attendance as estimated in 2014.
- X_8 = percentage of individuals adherent to Protestant Christianity in the state population in 2014

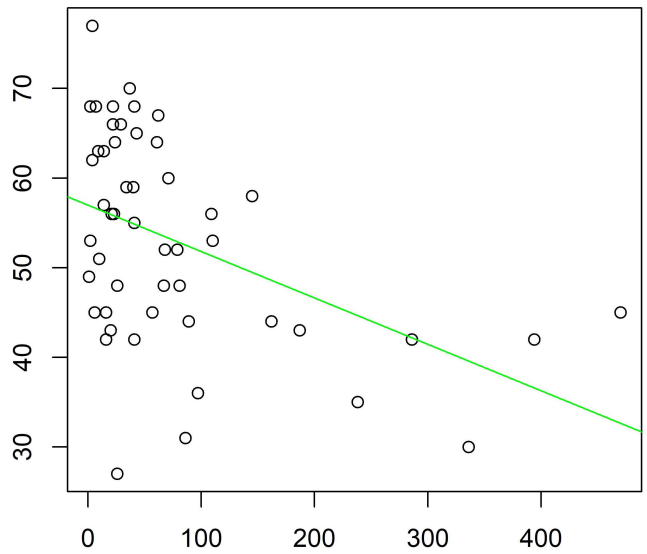
Exploratory data analysis

- Linear model

$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_8 X_{i8} + e_i, \quad i = 1, \ldots, n$$

- In the matrix notation $Y = X\beta + e$
- $R^2 = 0.78$
- Breusch-Pagan test $p = 0.766$, heteroscedasticity is not an issue here
- Controversial result: Y directly proportional to the population density, but $\beta_3 > 0$
- Large population density in the urban states of the New England

The election results against the population density:



Model building

- Standard backward selection by t-tests
- Submodel with 4 relevant predictors
  - X_1 = percentage of African American population
  - X_2 = percentage of Hispanic and Latino population
  - X_5 = percentage of individuals with a bachelor's or higher degree
  - X_7 = weekly church attendance
- Akaike information criterion finds the same submodel

Full model:

- $R^2 = 0.78$
- The predictors are largely correlated (largest $|r|$ about 0.70)
- Condition number of the matrix of predictors very high (4986.7)
- Serious problem with multicollinearity!

Submodel:

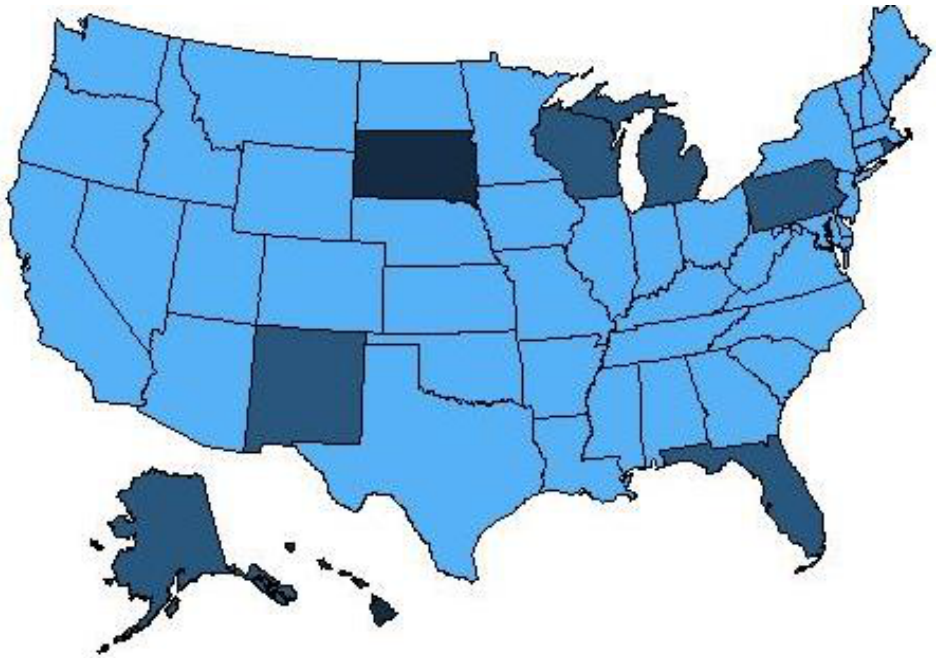
- $R^2 = 0.75$
- Again, the predictors are correlated (largest $|r|$ about 0.58)
- Condition number of the matrix of predictors much improved (57.1)

Outlier detection

Full model:

- Predictions much improved after deleting two severe outliers
  - South Dakota (the Democratic candidate withdrew before the election)
  - Hawaii (very specific demographic structure)
- Wrong prediction of the winner for 6 states shown in the figure
  - 4 of them are in fact the "swing states"

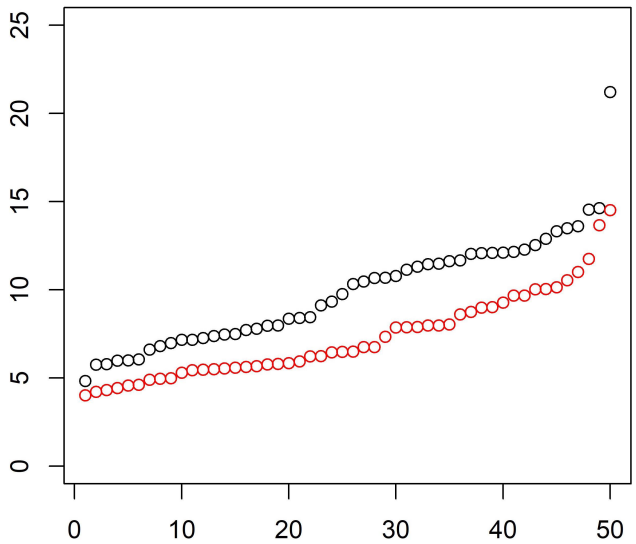
Submodel: analogous results



Confidence intervals for Y

- Crucial criterion of reliability
- Assuming normal errors and homoscedasticity
- The width of the confidence intervals is proportional to the leverage scores, i.e. diagonal elements of $X(X^T X)^{-1} X^T$
- Removing multicollinearity is thus beneficial
- Utah
  - Outlying in several predictors (ethnic minorities, religious demographics)
  - The widest confidence intervals
  - The prediction is very unreliable
- Narrowest confidence intervals: Ohio and Iowa
  - The demographic structure is very typical

Sorted lengths of confidence intervals for the full model and for the submodel:



Local sensitivity

- We add small perturbations to the predictors (but not Y)
- MSE considered in a 5-fold cross-validation
- Full model and the submodel turn out to have a sufficient local robustness

No. of states	No. of predictors	R^2	MSE
Raw data			
50	8	0.78	29.9
50	4	0.75	32.8
48	8	0.84	18.9
48	4	0.82	21.0
Local modification with normal distribution			
50	8	0.77	31.1
50	4	0.74	35.3
48	8	0.79	23.9
48	4	0.77	27.4
Local modification with uniform distribution			
50	8	0.76	31.8
50	4	0.72	37.5
48	8	0.80	23.2
48	4	0.77	27.1

Conclusions

- The model is meaningful with a quite large R^2
- The effect of demographic predictors on the popular vote has been known
- Our work is focused on a study of reliability
- The **submodel** with 4 predictors is **more reliable** than the full model
- Key aspects: dimensionality reduction, model choice
- Limitations of the study
  - A simple set of predictors on the state-wide level
  - Some outlying states not explained well by the predictors
  - Robust statistics not used here

Criterion	Which model preferable
Multicollinearity	Submodel
Outlier detection	-
Confidence intervals	Submodel
Local sensitivity	-