Emory University

The Multicollinearity of Linear Regression Model about Reagan's Feeling in 1980 Election

Qingyuan Zhang

Courtney Brown

POLS 301W 4:00 TTH

Assignment 6

18 November, 2015

The presidential election in 1980 between Reagan and Carter is a contest of personalities for both of the candidates. Eventually, challenger Reagan gain a dominant victory against Carter with his optimisms to the future. During the whole election, beliefs about the Reagan's personalities and capabilities has been thoroughly expressed in the feeling about the candidates. Voters' attitudes towards Reagan, which depends on plenty of factors might directly affect the final consequence of the election. In order to investigate the factors that determine the people's feeling about Reagan, I construct a multiple variable linear regression model to analyze people's feeling about Reagan.

My data to conduct the model is from the 1980 national elections panel study (NES). In the dataset, the feelings for Ronald Reagan has been divided into 4 waves, which the data is collected on January, July, September, and November respectively. I initially use the third wave voter's attitude as my dependent variable and choose nine independent variables in the regression model. The unit of our analysis is each individual that participate the survey about their feeling for Reagan.

To improve my preliminary model, I decided to test whether my choices of independent variables are good. Multicollinearity, which is one of the problem of dependent variable, occurs because of the correlation between my dependent variables. I test the multicollinearity of all my dependent variables, in which are the product of other independent variable, with variance inflation factors. Product of independent variables are supposed to have multicollinearity problem since it prone to correlated to their factorized variables. As shown in table 1, the multicollinearity for education, age, race, the product of education and age, and product of education and race have relative small variance inflation factors. However, compare with these dependent variables, income, product of education and income, product of race and income, and

product of education, race and income have very large variance inflation factors, which indicates the serious multicollinearity of the variables. However, the ordinary least square estimates are unbiased. As a result, to further test my original multiple variable linear regression model, I also conduct a ridge regression with all my dependent variables. The result behaves basic same as the variance inflation factors test. However, according to the figure 1, the product of education and income turn from negative to positive and the product of education and race turn from positive to negative, when we add some biased to our regression. But when we see the ridge regression result of income, we see that it has a decrease when we ass bias to the model, but the sign is always the same, so we decided to keep income as dependent variable but eliminate the the two products. To ameliorate my linear model of Regan's feeling, I decided to keep education, age, race, and income as my dependent variables but delete all the products.

(Insert table 1 and figure 1 here.)

As discussed above, the product socioeconomic variables are all eliminate from my model. To time two variables, the product variable contains the information of both variables. As a result, it is highly possible that the products are correlated to either of the factorized variables. To delete all those variables from my model, it makes my regression change on a small amount in response to the small change of the data. To explain the high variance inflation factor of income, I believe it might be its correlation with education. High education level people have a better change to gain high salaries. A regression with both income and education as dependent variables may return a big variable inflation factor.

According to my test result, I modify my original linear regression about Reagan's feeling. Education, age, race and income are the independent variables I kept in my model. To test the regression again, I did the variable inflation factor test and ridge regression. According to table 2, all the dependent variable returns a variable inflation factor less the two. The ridge regression, as is seen in figure 2, also turn to be stable. Based on my constructive test result, my new linear regression model is far better than before dealing with multicollinearity.

(Insert table 2 and figure 2 here.)

My model of people's feeling toward Reagan has initially suffer from the multicollinearity. I check all the initial dependent variables and only keep the variables that return a small variable inflation factor and stable result in ridge regression. The improved model is more convincible when there are biased or change in my original data. The better regression allows me to generate more inference about the political insight of 1980 election.

Table 1: Variance inflation factor for Reagan model of initial model

Independent Variable	Variance inflation factor
Education	60.61634
Age	29.40264
Race	50.35484
Income	223.32737
Education: age	29.81668
Education: income	352.29669
Education: race	87.24716
Race: income	205.74697
Education: race: income	312.29987

Table 2: Variance inflation factor for Reagan model of improved model

Independent Variable	Variance inflation factor
Education	1.225293
Age	1.114117
Race	1.049579
Income	1.245437

Figure 1: Ridge regression for Reagan model of initial model

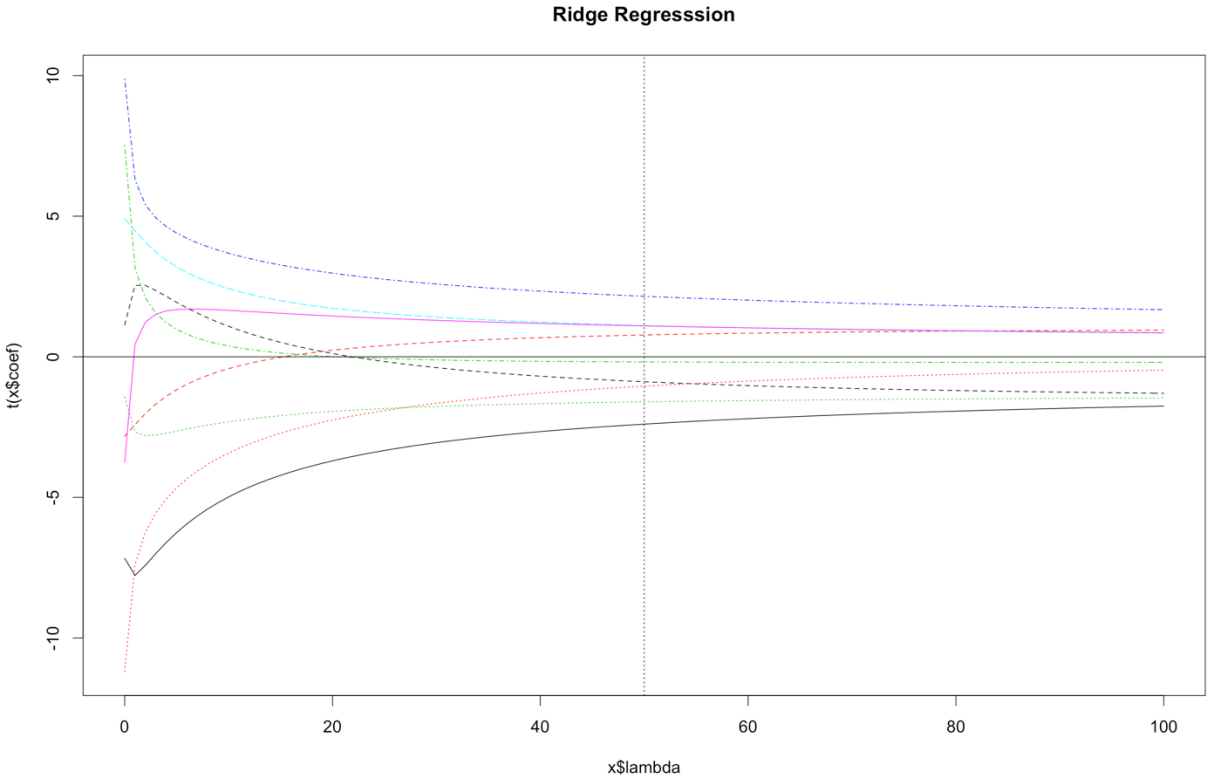

Figure 2: Ridge regression for Reagan model of improved model