Data Science Practical

Case 3 - Academic year 2021-2022

Part A: Winning elections by spending more on campaigns?

Suppose you work for a political party and campaign time is coming up. You are wondering whether you can attract more votes by increasing campaign expenditures. To see whether such a relationship really exists, you find a data set from the United States on election outcomes and corresponding expenditures for a large number of two-party races for the U.S. House of Representatives in 1988. You will focus on candidate A.

A simple linear regression model can be used to investigate the matter above. However, you wonder whether you should take just the expenditures of A into account or rather a measure of how much the political party spends *relative to the other*. Would it be possible to include both? Or would this lead to a problem of collinearity? Besides, you notice that the data set contains variables in levels and in log-levels (which ones do you prefer?). You build a more elaborate model including other factors that you believe explain the voting results. Using a set of measures, you hope to show that this new model "outperforms" the simple linear regression model. You hypothesize that the effect of an increase in expenditures by party A might be completely offset by a similar increase in B's expenditures and test this.

You remember that it is possible to account for nonlinear effects in a linear regression model. In particular, do increases in expenditures lead to more votes in a linear fashion or does this effect wear off? Would it make sense to include interaction terms in the model? If yes, what would this mean? And in case you include any of such terms, could you still calculate *partial effects* for party A? In every step of the analysis, you make sure that you explain everything in solid statistical and economic terms, since you are planning on sharing your interesting results. Hence, your story should convince both experts and laymen: a careful explanation of (the interpretation of) estimation results is thus important.

Since your data set was relatively small in terms of number of variables, you decide to have a look at another data set that contains voting outcomes in 1990 for incumbents that were elected in 1988. In particular, you wonder whether past achievements guarantee success in the future. That is, when voters believe the quality of the candidate is high in 1988, does that lead to more votes in 1990? However, "quality of a candidate" is not directly available in the data set. Could you think of a variable that might be used as a proxy? Instead of investigating the second data set, you may also choose the following alternative. Predicting election outcomes have been studied widely in econometrics. Often, the focus was different than the one in this case. That is, the researchers were mostly interested in identifying the key variables explaining why people vote for a specific party (e.g. high economic growth and low inflation of incumbent party). More information can be found on `https://fairmodel.econ.yale.edu/`. If you choose this last option, let me know, and I provide you some extra assistance!

Practical Information

This part of the case makes use of the data set in **data3a.xls** (and potentially **data3b.xls**). The goal is to apply a set of (modeling) techniques and to correctly interpret their outcomes. It is important that you are aware of the strengths and weaknesses of the methods you use. The story above should help you to gradually build up an appropriate regression model. Hence, it is a good idea to follow these recommendations closely, but note that creativity from your side is also rewarded. However, make sure you can always argue in favor of the decisions you take. The mandatory part of this case entails that you should use both **linear regression** and ***K*-nearest neighbor regression**.

Part B: Monte Carlo Study

Monte Carlo simulation studies are often employed to investigate the finite sample behavior of e.g. estimators and tests. In this study, we want to examine the consequences of heteroskedasticity on the OLS estimator. More specifically, suppose the true data generating process (DGP) is given by

$$Y = \beta_0 + \beta_1 X_1 + \varepsilon, \tag{1}$$

where ε exhibits heteroskedasticity. In reality, the DGP is not known to the econometrician. However, by simulating the data according to (1) we can study the effect of a non-constant error variance on the OLS estimator. Before executing the simulation study, explain what you expect will happen and why.

Practical Information

For this part of the case you are given a lot of freedom. Decide upon the setup of your simulation study. It could be useful to investigate the finite sample distribution of e.g. β_1 whenever the model is free of heteroskedasticity. Whenever you have results on this, you can compare to the case of non-constant error variance. Think about different measures to investigate these issues and decide on how you can effectively present the results of your simulation study.