E401/M518: Video Presentation

Factorization: Principal Component Regression

Fall 2023

November 19 2023

Please work on this video presentation in group of three students. All presentation assignments are based on real-world data and are similar to what you might encounter in your future career. The main purpose of this challenge is not on the application, but on getting you some practice with learning new empirical methods that we have not discussed in the lecture. If you pursue a career as an empirical economist or data scientist, you will have to be able to familiarize yourself with newly developed methods constantly. This includes learning enough about the methods to apply it to your day-to-day work and assessing in which circumstances the method is appropriate as well as what its limitations are. Nevertheless, whatever you present should make economic sense and if it doesn't you should think about what could be going wrong with the data and/or your analysis. For this assignment, I have tried to provide you with reasonably clean data, so that you should not have to spend a lot of time on data preprocessing tasks. However, I haven't checked every detail. Therefore, you should be prepared to have to do minor adjustments to the data if you run into any issues with your analysis. I strongly encourage you to come to my office hour to discuss any roadblocks as well as your overall plan for your video presentation ideally a few weeks before the assignment due date. There is always a risk that you won't be able to understand every detail of the method or the data. As long as you are able to clearly document what you tried and have some conjecture/explanation for why you get the results you get, this is totally fine. It's very likely that you will be in similar situations regularly when taking a job as an empirical economist or data scientist.

You are expected to upload a video presentation of roughly 25-30 minutes. Think of this presentation as one you would give to your boss or at a board meeting of a company or policy institution that hired you as a quantitative analyst. The main focus of the video presentation is to communicate effectively the essence of a new empirical method and illustrate how it works in an application. Anybody who has take none or two econometrics classes and the undergraduate level should be able to follow your presentation. The students refereeing your video should think of themselves as board members who attend your presentation. When you write up your referee reports, you are strongly encouraged to provide critical questions and constructive comments about the presentation. That is, you should provide a brief assessment of how helpful you found the presentation in learning the discussed methods, potentially provide corrections, and ask any remaining questions that you may have either regarding the

method or the discussed application.

Each video presentation will likely have a similar structure and should contain the following elements:

1. Discussion of the method: (1) What is the main idea of the method, (2) which problems does the method solve, (3) how does the method solve the problems, (4) what are the conceptual steps involved when implementing the method, (5) what decisions do you have to make when applying the method to a specific data set, (6) what are some common pitfalls ad when would you not want to apply the method.
2. A brief discussion of the data: Compared to the challenge presentations in class, this part should be very brief. Nevertheless, remember that not everybody who watches the presentation may have looked at the data before.
3. Discussion of the application: (1) What's the big picture business or policy question that you are trying to address? As usual, your viewers probably have not read the assignment questions in advance, (3) discussion of your empirical results, policy implications, and potential caveats and suggestions for further steps.

As with the in-class presentations, this is not a presentation class, so don't invest in fancy video production elements! A fairly bare-bones recording in which you walk the viewers through a couple of slides - you should not need more than 10 for the method explanation- and an RScript that generates all your results as you click through it, and explain it, is totally fine.

Main Techniques

Principal Component Regression (PCR)

Key references

In order to work on this assignment, you will have to be familiar with principal component analysis; therefore, it might be a good idea to talk to the group that works on the PCA assignment. A good starting point to learn about PCR is the corresponding chapter in our ISLR textbook.

Data

In this assignment, you will work with survey response data from a consumer focus group who viewed and rated several TV show pilots. The data contains 6241 viewer-show observations with 20 questions each and 40 shows were rated. There are two groups of questions: Q1: *This show makes me feel ...* and Q2: *I find this show feels...* The survey responses are contained in `nbc_pilotsurvey.csv`. The file `nbc_showdetails.csv` contains information on the ratings of each show:

1. `Gross ratings points (GRP)` - classic measure in the broadcast industry of how popular a show is.
2. `PE (projected engagement)` - a more subtle measure of how attentive an audience is to a show.

In addition, the file contains genre information and the duration of a typical episode.

Context and policy question

This assignment might be particularly interesting if you are into US TV shows. In this assignment, you will build a model to predict the popularity of a show based on the pilot survey results.

1. Briefly discuss whether the data you are working with fits the criteria of small or big data as discussed in class.
2. First, aggregate the survey results by computing the average response for each show and question. Afterwards, reduce the dimensionality of this summary matrix by running a PCA. Look at the implied factor loadings and try to interpret them, especially the first one.
3. Now visualize each TV show in the PC1-PC2 space. Briefly, interpreted some aspects of the graph. Are you surprised by the results?
4. Now, build a model to explain audience engagement as a function of the principal components. Discuss in detail what decisions you have to make when you set up your model, in particular, the choice of how many principal components to include. Discuss two ways of how you could choose the number of principal components. Estimate both models and compare the results.
5. Finally, compare your PCR to a LASSO model estimated on the original survey statistics, i.e., the 20 questions. Discuss in which settings LASSO on the original data would be your preferred model. When do you think a PCR model makes more sense? Are you surprised by your findings in this application?
6. Optional: If you have time, briefly discuss the main disadvantage of PCR and how partial least squares (PLS) could improve on PCR.

Hints:

1. There are several ways to do PCA in R. In my experience, the most efficient is the command `prcomp`.