Data Science Practical
General instructions - Academic year 2024-2025

Goal

In this course, you obtained a real-life data set and – simply said – the goal is to analyze it the best way you can. The short case descriptions on Canvas indicate the main topic of your research and help to choose your variables of interest. As an example, imagine that you are assigned a case that considers investigating the relationship between sales and advertisements. This task may be relatively straightforward if the variables at your disposal are well-defined. However, it becomes more difficult if your data set consists of different types of advertisements, incomplete measures of advertisements or maybe not even a clear measure of advertisement at all. In these cases, you have to be creative and think of clever solutions.

Another complication might occur in deciding on the auxiliary variables. If your data set consists of many variables, it might be difficult to decide which ones to include in your model and which ones not. Base your choice on different arguments. Let's provide some examples:

- **Use logic**. As an example, suppose your dependent variable y is binary and measures whether a person is in the labor force or not. Then it might not be a good idea to regress on the variable wage, as there is a clear endogeneity issue: being in the labor force determines whether you have a wage or not. Thus, not all variables in your data set might be suitable to be used as regressors.
- **Use related literature**. There might be articles that study a problem similar to yours. In this situation, you could refer to this study to motivate your choice of variables. It is often the case that the articles base their decision on *economic* and *financial* theories.
- **Use statistical arguments**. You have studied a variety of tools that can help you to discriminate to between models. In many cases, empirical researchers base their modeling decisions on a combination of different measures.

Make sure that you use convincing and theoretically sound arguments to convince the reader of the report of your modeling choices.

Action plan

For this project, it might be advisable to take the following steps:

- **Prepare your data**.
  - Have a good look at your data by making plots.
  - In case of outliers, you should consider deleting them.
  - Does your data show certain nonlinear patterns? Maybe you can transform the data (e.g. using a log-transformation) to reduce the degree of nonlinearity.
  - Compute descriptive statistics (mean, variance, minimum, maximum, etc.) and/or frequencies to summarize the main properties of your data.
  - Compute correlations between variables (for which it is applicable) to get a preliminary idea of how they relate bivariately to each other.
- **Divide your data in training and test set**.
  - Chapters 2 and 3 of the books "The Elements of Statistical Learning" and "An Introduction to Statistical Learning" (available at `www.statlearning.com`) provide good explanations on these concepts. This is generally the case for this project, as they cover the linear regression model and K-nearest neighbor regression in quite some detail.
- **Inference: linear regression**.
  - Simple linear regression models might be a good starting point, as results can be nicely visualized. It might give you some insights on variables that are promising to include in your model.
  - You can extend your model gradually. Consider the possibility to include categorical variables (you can use them as factor variables: they allow for considering subsets of the sample), interaction terms (variables that "work together" and create synergies in this way) or nonlinear transformations of regressors (e.g. if one believes that an effect is not linear, but wears off at some point, you could consider including both x and x^2 in your model).
  - You can perform hypothesis tests to learn more about relationships between variables. Note that you have studied tests for single and joint significance. You have also seen confidence intervals.
  - Check and/or test the underlying assumptions of your model such that you can make possible adjustments and to comment on the reliability of your analysis. Note that you cannot always 'solve' the problem of assumptions not being met. If you are dealing with such a situation, it might be interesting to investigate it in more detail in the Monte Carlo simulation study.

- **Prediction: linear regression, *K*-nearest neighbor regression, regression trees**.
  - Select some linear regression models to make predictions.
  - You can decide on the variables to include for *K*-nearest neighbor regression. It can be the same variables as you included in the linear regression model, but you can also choose to run multiple specifications.
  - It is important to decide on the number of neighbors *K* (also called a hyper-parameter in machine learning) and appropriate distance measure(s) to determine 'closeness between data points'. Chapter 5 of "An Introduction to Statistical Learning" explains how cross-validation can be used to determine *K*, but there are also other methods to do so.
  - For regression trees, specific splits have to be made in the available features. Once again, you might want to use cross-validation. For more information on regression trees, check out Chapter 8 of "An Introduction to Statistical Learning".
  - *K*-nearest neighbor regression, linear regression and regression trees can be used to perform predictions (and possibly prediction intervals). You can compare the predictions of different specifications to see whether you would use the same model for inference and prediction purposes. The prediction performance of the specifications can be compared using metrics such as the *mean squared (prediction) error* or *mean absolute (prediction) error*.
- **Run a simulation study**.
  - In order to have a coherent scientific report, it is nice if you run a simulation study that considers a problem you face in your analysis. You are asked to compare the performance of the linear regression model, *K*-nearest neighbor and decision trees (select at least 2) in terms of prediction. You can follow a similar approach as in Section 3.5 of the book "An Introduction to Statistical Learning". However, note that you can extend their analysis to data having heteroskedasticity, outliers, possible strong (imperfect) multicollinearity. You can also think about other investigations: what happens when you forget to scale your data in the *K*-nearest neighbor approach? Or what are the consequences of using different distance measures? There are various possibilities. Clearly, you do not have to implement them all, but it could be interesting to choose a scenario that fits your empirical study well.
  - It is very important that you are clear on the data-generating processes you consider. You can study cases where the assumption is not violated, only slightly violated or violated very badly. Hence, it is crucial to think carefully about the setup of your simulation study. Approach us if you need help with it.

Report

When we discuss academic writing, we will also discuss what a good structure for your report could look like. The idea is that you write a scientific report that comes close to an academic article. Therefore, it is not only important to write mathematical equations, but also to interpret your estimation results. This means that you do not only look at whether coefficients are significantly different from zero, but also what the results mean in terms that non-econometricians can understand. Do your results have important implications? Could it, for example, lead to policy changes?