The Process of Data Analysis

Assume that we have obtained the data randomly sampled from the population. Now we are going to establish a linear model to describe the pattern of correlation or dependency. We specify a 6-step process as follows:

- **Step 1. The narrative** *(problem and variable specification)*

  Use plain words to **describe** the problem, including the background and the variables to be considered, etc.

  You should think thoroughly about the nature of the problem: whether it is regression problem or a classification problem? Or maybe a clustering problem? (In our course, we may mainly focus on the first two types of problems.) You may feel free to imagine and try all possible methods to dig into the data at hand.

  After you are clear about the nature of the problem and the details of the available data, you can choose your explanatory variables and response variables. Generally, response variables are what you would like to 'predict'. However, in an open setting, you may defer the decision until after the second step.

- **Step 2. EDA** *(Explorative Data Analysis)*

  Explorative data analysis (EDA) is generally the first step in exploring the data. Various methods may be used, most of which are *intuitive graphical methods*. For example:

  - *Histograms*: histograms gives useful information about how a random variable is distributed, whether it is symmetric or whether it is heavy tailed, etc. After checking the histograms, sometimes you may feel it necessary to pre-process the data. For example, you may perform log transformation to improve the property of the distribution.
  - *Pairwise scatter plot*: scatter plots gives information about correlation among different variables.

  You may also check some **descriptive statistics**, including sample mean, sample variance, skewness, and the correlation matrix, etc,

  After a first look into the data, you may run some simple linear regression models to verify your findings.

- **Step 3. Model specification** *(parameter estimation & model comparision)*

  Officially establish a model to describe the data. You should first make some **model assumptions**. For example, when you want to use linear regression to model the problem, you should follow the assumptions like normality and homoscedasticity. Then, you can formulate the problem into the form of mathematical model, and perform parameter estimation afterwards.

  You may establish more than one models, and compare them afterwards. After choosing a set of specific criteria, you can select the `best` model. During this process, you can either repeat the modeling-and-testing process, or you may try using automatic model selection.

Notice that, there may be a 'best' model according to some criteria, but remember that 'all models are wrong, but some are useful'.

- **Step 4. Diagnostics and Remedies** *(Check model assumptions and make ammends)*

  However, the seemingly complete process in the previous step is not enough. It's critical to perform model **diagnostics**. In plain, diagnostics is intended to check whether there is violation of model assumptions. After diagnostics, you may need to carry out some modifications. For example, you may have to give up the original model and try ridge regression when heteroscedasticity is significant in the residuals. We have to make sure that no assumption is violated.

- **Step 5. Inference and prediction**

  After the previous steps, we have a reliable linear model at hand. Now we are going to make use of this model. You can make inference of mean value or future observations. (These two inferences are different in that the latter one considers the random noise.) What is amazing about regression models is that you can have a nice estimation at some unobserved places in the sample space, which may be referred to as 'generalization' power. However, you should still be cautious about the model assumptions, as the model will lose efficacy beyond the scope of the hypothesis. Also, you may try to visualize your results.

- **Step 6. Evaluation and modification**

  After putting the model into practical use, you and your coworkers can evaluate and modify the model, which will be helpful for future work.

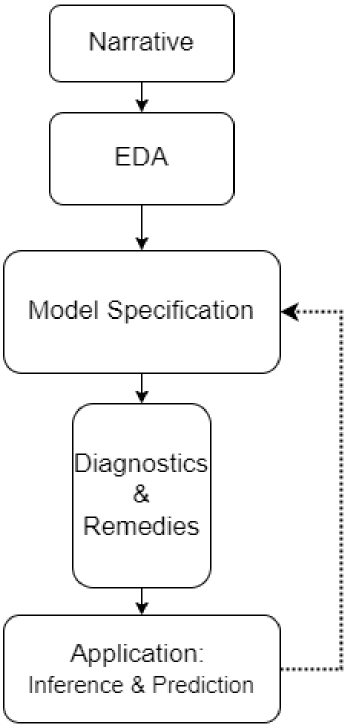

figure 1. The Process of Data Analysis