Statistics 352

Course Project

The goal of the course project is to find a dataset (ideally, one that is of interest to you) and analyze it using one (or more) of the nonparametric methods we have discussed in the course.

Project Topic

The project can be based on either nonparametric density estimation or nonparametric regression. Below I give some examples of the types of analysis that would make a good project. However, it is important that the methods used are appropriate for the data; in particular, you do not necessarily need to include all of the components I describe.

1. Nonparametric density estimation

Projects based on density estimation typically fall into one of three categories:

- estimation of the density of a random variable together with investigation of the properties of the estimate and/or a comparison to a "standard" density for the data
- a comparison of the density functions of a random variable for two groups
- the use of density estimation for classification.

Note that some projects might combine features of more than one of these categories, if appropriate for the data. In all cases, the distribution of the "response" variable should be continuous, in the sense that it is appropriate to model it as a continuous random variable with an unknown density function.

Further details are given below.

Estimation of a density

Your data should consist of independent observations of a random variable Y such that the density function $p(\cdot)$ of Y is of interest. The analysis might include the following components.

- An estimate of $p(\cdot)$ with a discussion of the selection of the smoothing parameter
- A summary of the properties of $\hat{p}(\cdot)$, possibly including quantities obtained by numerical integration

- A comparison of $\hat{p}(\cdot)$ to a standard density based on a test of the hypothesis $p = p_0$ for a given value of p_0. The standard density should be one that is interesting in the context of the data, e.g., a normal distribution for measurement data or an exponential distribution for failure data.

- You could also include calculation of statistics that are not directly based on the density estimate, using the bootstrap method to find the standard error of the statistic. For this type of calculation to be approperiate, the statistic(s) should be one that has a useful interpretation for the data being analyzed.

Comparison of two groups

The goal for this option is to compare the densities of a response variable for two groups and to use the value of the response to predict the group to which the response belongs.

Your data should consist of a response variable Y for a set of "subjects", each of which belongs to one of two groups, which can assumed to be known, in the sense that there is no uncertainty regarding the group to which a subject belongs. The analysis might include the following components.

- Estimate the density function of Y for each of the two groups (call them group 1 and group 2)

- Summarize informally the differences between the two estimates

- Test of the hypothesis that the true density functions for the two groups are identical.

- If the densities appear to be different, test the hypothesis that the shapes of the true density function for the two groups are identical (if appropriate)

- Summarize the comparison of the two densities in the context of the data.

Classification

The goal for this option is to compare the densities of a response variable for two groups and to use the value of the response to predict the group to which the response belongs. The data can be either univariate or bivariate.

- Estimate the density function of the response for each of the two groups (call them group 1 and group 2).

- Compare the estimates and, for univariate data, test the hypothesis that the true density functions for the two groups are identical.

- Suppose that, on the basis of an observation $Y = y$, we wish to predict the group to which the observation belongs. Estimate the conditional probability that an observation is from group 1 given that $Y = y$, as a function of y and plot the results.
- Interpret the results in the context of the data. One way to do this is to give the estimates of this conditional probability for a few values of y.

2. Nonparametric regression

Projects based on nonparametric regression typically fall into one of two categories:

- Estimate a nonparametric regression to a response variable and a predictor variable and use that estimate to summarize the relationship between the variables. Both variables should be continuous.
- Estimate the parameters of a semiparametric regression model relating a continuous response variable to a continuous "nonparametric" predictor and a "parametric" predictor which does not need to be continuous.

Further details follow.

Estimation of a nonparametric regression function

Your data should consist of observations on a response variable Y and a predictor variable X; of interest is the regression function $m(\cdot)$ given by $m(x) = \mathrm{E}(Y|X = x)$. The analysis might include the following components.

- Use kernel estimation to estimate $m(\cdot)$; include a discussion of the selection of the smoothing parameter
- Summarize of the properties of $\hat{m}(\cdot)$ in the context of the data; this might include providing estimates of $m(x)$ for meaningful values of x
- Find the degrees-of-freedom corresponding to the estimate $\hat{m}(\cdot)$ and estimate σ, the error standard deviation.
- Test the hypothesis tht $m(\cdot)$ is constant (i.e., there is "no effect") or that $m(x)$ is linear in x (or both), if appropriate for the data.
- Compare the estimate $\hat{m}(\cdot)$ to the estimate that would be obtained using a polynomial regression model, if a low-degree polynomial regression model might be appropriate for the data.
- Summarize any conclusions regarding the relationship between Y and X that result from the nonparametric regression analysis. Any such conclusions should be discussed in the context of the data.

Semiparametric regression

The goal for this option is to analyze data using a semiparametric regression model.

Your data should consist of a continuous response variable Y along with a two predictors: Z, which is continuous, and X which can be continuous, discrete, or categorical. The relationship between Y and X will be modeled parametrically; hence, this relationship should be approximately linear (unless X is categorical). The relationship between Y and Z will be modeled nonparametrically.

- Estimate the regression function $m(z) = E(Y|Z = z)$ using a local linear kernel estimate.
- Find the degrees-of-freedom of your estimate and estimate the error variance.
- Test the hypothesis that $m(\cdot)$ is a linear function.
- Estimate the parameters of the semiparametric regression model

$$Y = \beta X + m(Z) + \epsilon.$$

  Provide an estimate of β along with its standard error.

- Interpret the results in the context of the data.

Data

You will need to find the data to use in your analysis. If possible, your data should address a question that you find interesting.

The only requirement is that the data include a sufficient number of observations for nonparametric estimation to be reasonably accurate. For the density estimation option, there should be at least 25 observations in each group; for the regression option, there should be at least 25 observations. In both cases, larger sample sizes are preferable.

For the density estimation option, most datasets that have been collected for the purpose of comparing two groups will be appropriate.

For the regression option, many datasets used to measure the effect of one predictor variable, while controlling for another (e.g., analysis of covariance), will be appropriate. Nonparametric methods are most useful when the relationship between Y and Z is nonlinear; however, it is fine if for your data the relationship is approximately linear.

Your data can be data that you have collected, perhaps for some other purpose or as part of another analysis or another course; alternatively, there are a number of sources where you can look for data that are of interest to you. These include

- `https://vincentarelbundock.github.io/Rdatasets/datasets.html` contains a list of the datasets that are available in a wide range of R packges.
- `https://archive.ics.uci.edu/ml/datasets.html`, the University of California, Irvine Machine Learning Repository.

- `http://lib.stat.cmu.edu/datasets/`, Statlib.
- The book *A Handbook of Small Data Sets*, by Hand *et al.*, which contains descriptions of about 500 datasets, has been placed on reserve in the main library. The book also contains the corresponding datasets, although these are easy to find (and download) on the internet.
- For those interested in sports, the websites `https://www.baseball-reference.com/`, `https://www.pro-football-reference.com/`, `https://www.basketball-reference.com/`, and `https://www.hockey-reference.com/` contain extensive data on the players and teams of the respective sports.

Your Report

Summarize your results in a brief report. It should include

- A description of the data you used, along with its source; if you have collected the data yourself, include a description of how the data were obtained
- The goals of the analysis, stated in the context of the data. For instance, if you are estimating a nonparametric regression function relating blood pressure to heart rate (for example), you should state the goal in terms of what you hope to learn about blood pressure and heart rate.
- A summary of the results and a brief description of your analysis. It's fine to present some R output to support a statement in that description; however, you should not submit all of the unedited output from the R functions used. In particular, it is important to interpret any numerical results in the context of the data.
- Include any plots and figures needed to understand your results. For instance, when estimating a density or regression function, include a plot of the function estimate.

Be as concise as possible; the goal is to provide a well-written and informative summary of your analysis.

Submitting Your Report

The project is due at 3 pm on Tuesday, June 12. Please submit it on Canvas.