Course Project

Check important dates available on Canvas

1 Objective

There are three main goals our course project aims to achieve:

1. The primary goal is to have you practice theories and concepts learned in class in application to a real life data using R as your primary data analysis software.

2. The secondary goal is to practice writing full-scale data analysis reports, similar to the ones you will be preparing in our other classes or as part of your future job as a data analyst.

3. The final goal is to have you prepare and deliver a short presentation of your findings, as well as participate in the discussion of findings delivered by your classmates.

2 Data

You have two options with regards to the source of data for your project. The first option is to use one of the datasets described below. The second option is to find data elsewhere, e.g. on Kaggle. If you decide to go with the latter, you need to (a) make sure that your dataset contains at least 1000 observations with at least 10 explanatory variables and (b) confirm your chosen dataset with me.

The following four datasets are available as .RData files[1] through Canvas:

- **Iowa liquor sales data.** This is a merged dataset based on sales of liquors in Iowa along with Iowa's demographic and economic data available through American Community Survey. The dataset contains average annual sales (in dollars and liters) of each liquor category across all Iowa's zipcodes across the 5-year period of 2012-2016.

[1].RData files can be opened directly with RStudio and the data in them is good-to-go for estimation.

- **Basketball salaries.** The dataset comes from basketball-reference.com and contains players' salaries alongside a collection of various annual performance metrics for each player.
- **Baseball salaries.** This dataset is part of the Lahman Baseball Database and contains players' salaries alongside a collection of various annual performance metrics for each player.

Each of the four datasets above contains both the data itself as R's dataframe object (e.g. `iowa`), and a description of the variables in form of another dataframe with the name ending in ".desc", e.g. `iowa.desc`.

3 Research agenda

While the actual research questions will be determined by the dataset, there is an overarching agenda that every project is supposed to follow. Namely, each group must analyze an outcome variable of interest as a function of several explanatory variables (features) either in a regression or a classification model, and the goal is to come up with the best model for two different scenarios: causal analysis (inference) and pure prediction.

4 Instructions

1. **Important:** make sure to watch the video on Canvas explaining certain features of the project in greater detail.
2. For both inference and prediction scenarios your baseline model should be multiple linear regression. From there you can go into more complex models such as non-linear regression, trees, penalized models, neural nets and so on. Naturally, your inference models should generally be not too complicated, as model complexity is inversely related to model's interpretability. In most cases your best inference model will be some extension of baseline linear regression, e.g. with added squares and/or dummy variable interactions.
3. While it is up to you to choose what you think is the best metric for comparing models, you should always compute that metric based on train/test sample split, ideally via k-fold cross-validation.
4. Your first draft submission should contain:
   (a) the description of the origin and structure of you dataset
   (b) the names and nature (units of measurement) for all the variables
   (c) basic numerical summary for all the variables (min/median/mean/max)

(d) key visualizations such as histograms and/or scatter plots for most interesting and/or important variables

(e) a few paragraphs of text explaining your research agenda, i.e. what is your main outcome variable of interest, what are the other key variables that might explain it, and what kind of results you expect to find from your inference and/or prediction models.

5. Your second draft submission should contain:

   (a) everything from your first draft submission with all the necessary changes as indicated by my feedback

   (b) estimation results from your basic inference model, causal interpretation of those results and a description of whatever else you plan to add to your inference model

   (c) estimation results from your basic prediction model or at the very least a description of what you plan to use as your prediction model

5 Project Paper Structure

The final version of the project paper should contain the following parts:

1. **Introduction.** Here you should explain the goals of the project and briefly outline any key expectations you might have (e.g. "Expecting higher sales of alcohol in areas with higher income per capita").

2. **Data summary.** This section should contain a short description of the nature of the dataset you are using, key numerical characteristics and some visualizations of key variables. At the minimum, it should include:

   - Name, type and units of measurement for all variables.
   - Basic numerical stats for each variable (min, max, median, mean).
   - Relevant/interesting visualizations for some variables (histogram and/or bar chart and/or box plot), if there are some noteworthy patterns (e.g. outliers).

3. **Research agenda.** Here you describe what kind of results you can expect for your inference model in terms of which variables are likely to have significant impact, whether certain subsamples of data (e.g. male vs female) are going to have notably different models via dummy interactions, and so on.

4. **Estimation results: inference.** In this part you should present your best inference model. Most likely it is going to be a multiple linear regression, and therefore you should provide your analysis with regards to:

- which variables were included and in what form (linear values, interactions, logs, polynomials)
- whether the model satisfies ZCM and/or normality assumption (the latter can be dropped if using large dataset)
- which of the variables have statistically significant effect on the outcome variable
- whether the sign and magnitude of statistically significant variables is consistent with common sense and economic reasoning

5. **Estimation results: prediction.** In this section you need to explain what different models you were using as candidates, what kind of metric/procedure you used to calculate the winner, and how the final best prediction model looks like. You also should compare that model with the inference model from previous section, if possible — e.g. are they of the same type, do they use the same set of variables, how much better pure prediction model is relative to pure inference model.

6. **Conclusion.** Here you should outline your key findings, as well as whether your expectations described in introduction were fulfilled.

7. **Appendix.** You should not crowd your main text with too many tables/charts, but instead put most of them into the appendix.