Contemporary Data Analysis: Survey and Best Practices

Project Assignment

Valentina Kuskova, PhD

Contents

Project objectives

In this course, which is a review course, you have been doing most of hands-on work on your own. Project assignment is no exception. There is one difference, however: you get to choose from very different projects, so you can do what you like the most. The main objective of the project assignment is to allow you to complete one analytic problem, from start to finish. Look through the "menu" of projects to see which one better fits your interests. Most importantly, have fun!

Submission requirements

Your project has to be submitted in a form of complete reports with all sections of the assignment completely identified.

1. If you are doing a model-based assignment, clearly state your variables and model setup
2. Include all code, calculations, and results. If they clutter your work, move them to the Appendix

Submission format

There are some requirements to the submission format as well. Please do the following:

1. Report may be done in any standard word-processing software, using Rmarkdown, but must be converted,if necessary, and submitted in the Word or PDF format.
    - Please note that for the 70-point grade, we are not asking you to spend time on formatting the document. It can be just the plain report, with the bare minimum requirements for readability. This is a course in CDA, not design, so if you are opting for the lower grade, you do not have to spend time on formatting.
    - Should you decide to make the report "better-looking" by highlighting the important items, getting proper headings for tables and figures, etc., you will be rewarded by 5 additional points towards the grade (as you saw in the table above).
2. All code for generating the results must be included either as part of the Rmarkdown document or in the appendix. If we do not see your code, we cannot evaluate the results. Failure to submit the code will result in a failing grade for the project.
3. Images, charts, tables and other descriptives should be included in the text in their corresponding section (not in the appendix or elsewhere). Please notice the word *should*. There is no penalty if you don't do this, but you might get a lower grade if we can't find your tables and figures.

Submission deadline and grading

- Submission deadline is clearly indicated in your course. Please utilize the Coursera system to submit your work.
- Projects submitted outside the Coursera submission system will not be accepted.
- Please allow up to 2 weeks for grade to be delivered to you, though we will try to provide you with feedback as soon as possible.

Data management project assignment (Lecture 2)

Part 1. Read the Huff's "How to lie with statistics" book

- Read the following two writings:
    1. Huff, D., 1993. How to lie with statistics. WW Norton & Company.
    2. King, G., 1986. How not to lie with statistics: Avoiding common mistakes in quantitative political science. American Journal of Political Science, pp.666-687.
- From the King's article, highlight:
    - What issues King covers that Huff does not?
    - Which of King's suggestions did you NOT understand? This list could serve as a starting point of items for you to learn in our Linear Models course, for example. We want to make sure you won't make these mistakes in your data work, and you have to know what NOT to do!

  - Provide ONE real-life example (at least, of course – you are welcome to find more) of data manipulation that could mislead the unsuspecting consumer
  - Hint: statistical analysis of elections is always a great source of such lies.
    * By the way, from all sides.
    * Also by the way, in all countries.
    * Also by the way, "lies" here are not the most obvious – you may not see them right away
  - Also look in media – it's all over the place
- Submit your explorations as a report no longer than 3 pages.

Part 2. Let's do some data management

- Using all of your skills and creativity, select one of the example files (01statess.xlsx, 01lottery, Enrollment2002) and convert it to a software-usable data format. (Hint: start with "Tidy data" reading we've had in second lecture)
- Doing it in Excel is fine, but feel free to explore other programs
- Submit the clean file of your data

Deliverables

To receive a grade of 70 points, you must complete part (1) of the project.

To receive a grade of 100 points, you must complete both part (1) and part(2)

Missing data assignment (Lecture 2)

This project is designed to help you better understand the effect of missing data on the outcome.

Take a complete dataset (with no missingness) of interest to you with two variables, x and y. Call this the "full data." By the way, the "Transportation Data," which you have in your Project folder, is an excellent candidate.

1. If you can, write a program in R (or any other software of your choice) to cause approximately half of the values of x to be missing.
2. Design this missingness mechanism to be at random but not completely at random; that is, the probability that x is missing should depend on y. Call this new dataset, with missingness in x, the "available data."
3. Perform the regression of x on y (that is, with y as predictor and x as outcome) using complete-case analysis (that is, using only the data for which both variables are observed) and show that it is consistent with the regression on the full data.
4. Perform the complete-case regression of y on x and show that it is not consistent with the corresponding regression on the full data.
5. Using just the available data, fit a model in R for x given y, and use this model to randomly impute the missing x data.
6. Perform the regression of y on x using this imputed dataset and compare to your results from (3).

Deliverables

To receive a grade of 70 points you must achieve the following:

- Complete the tasks (3), (4), and (6). In other words, you should be able to find a way to perform the three regressions - even if you have to create the required datasets manually.
- Submit our evaluation of the observed results (no more than 2 pages)

To receive additional 30 points, you must also do the following:

- Complete all other items of the assignment (programmatic data removal, ensuring missingness mechanism)
- Clearly indicate why you have chosen a certain approach for working with missing data
- Format the report so that it looks professional

Practical descriptive analytics assignment (Lecture 1, Lecture 3)

This project is designed to help you create meaningful descriptions. Use your imagination and creativity with the given data. In real life, the data given for this assignment are abound.

Datasets

There are two datasets for this study. They are very similar in composition, so it's a matter of your personal interest.

Automotive market data

This is the dataset for Russian auto market sales (total), starting from the year 2005.

- Data are monthly, from January 2005 until June 2018
- The following are the dataset variables:
  - Date - month and year
  - Total Sales - the number of new cars sold in this month
  - CPI - Consumer price index (inflation). Given in percentage points, so value 12.5 means 12.5%.
  - Oil - Prices for Brent oil, per barrel, in USD.
  - CCI - Consumer confidence index. Given in percentage points, so value 99.67382 means 99.67382%
  - BCI - Business confidence index. Given in percentage points, same as CPI.
  - Prime rate - interbank exchange rate. Given in percentage points, same as previous variables.

Ruble-USD exchange rate data

This is a dataset of exchange rates between Russian ruble and US dollar, with some additional financial data.

- Data are daily, from January 11, 2006 until March 12, 2020
- The following are the dataset variables:
  - t - time period (daily)
  - Date - actual date in format mm/dd/yyyy
  - Day - day of the date
  - Month - month of the date
  - Year - year of the date
  - RubPerUSD = exchange rate in rubles per US dollar

- SP500 - USD value of Standard&Poor's 500 financial index
- Gold, Silver - prices of gold and silver in USD per ounce
- Brent - prices of oil (Brent brand) in USD per barrel
- Cocoa - prices of cocoa beans, in USD per metric ton

Assignment

Choose any of the datasets. The assignment is identical, but do it on the dataset of your choice.

1. Create a table that will contain the following:

Variable Name	Quantitative or Qualitative	Measurement level	Appropriate charts

2. Fill the table with *all* of the variables in the dataset, classifying them appropriately. You will need to indicate whether variable is quantitative or qualitative, its measurement level (nominal, ordinal, interval, ratio, counts), and which charts (pie, bar, line, boxplot, etc.) are appropriate for this variable.
3. For any *three* variables in the dataset, create an appropriate chart. Feel free to be creative with colors, titles, etc.
4. For the same variables, calculate the appropriate measures of central tendency.
5. For the same variables, create a matrix of correlations.
6. Create at least one complex chart that combines information.
7. Create at least one complex chart that demonstrates results of the model. To do so, perform the following tasks:
   - Choose a dependent variable
   - Choose at least two independent variables
   - Build a linear regression model
   - Plot results of the model

Deliverables

To receive a grade of 70 points you must achieve the following:

- Complete the tasks 1-6.
- Comment on all statistics you calculate and all graphs you generate. What do you observe?
- Put results into one document with all the code and the output.

To receive additional 30 points, you must complete the following:

- Complete task 7
- Format the report so that it looks professional

Linear regression assignment (Lecture 4)

This project is designed to help you understand inferential statistics. You have to create a meaningful linear regression model.

Datasets

There are several datasets offered as part of the project assignment. Feel free to choose any of them (Transportation data, Auto market, etc.).

Assignment and deliverables

For any of the datasets of your choice, please do the following:

- To obtain the grade of 70 points, please do do the following:
  1. Choose a dependent variable. Explain your choice (no more than 0.5 pages)
  2. Choose *at least* three independent variables. Explain why you think these variables will affect your dependent variable (no more than 2 pages)
  3. Build a linear regression model. You can go through several iterations before your model satisfies all of the criteria of a good linear model:
     - The model itself is significant
     - Each of the independent variables (except for the intercept) are significant
  4. Once the model satisfies the criteria of good linear model, please describe the model:
     - Comment on the R-square and explain what it means
     - Comment on each of the variable coefficients and explain what each means
- To get an additional 30 points, please do the following:
  1. For the chosen variables, calculate measures of central tendency
  2. For the chosen variables, create a table of correlations
  3. For the chosen variables, create appropriate graphs and charts
  4. Once you build a satisfactory model, plot model results
  5. For each of the variables, comment on the t-statistics and p-value

Predictive model assignment (Lecture 5)

This project is designed to help you hone in on your predictive modeling skills. It follows Lecture 5 challenge very closely.

Data and setup

Dataset for this assignment, "CancerData.txt" is provided in your Project folder. Let's read it in and take a look at it:

```
cancerData <- read.csv("CancerData.txt", stringsAsFactors = FALSE)
```

It contains the following variables:

1. Sample code number (ID number)
2. Clump Thickness (1-10)
3. Uniformity of Cell Size (1-10)
4. Uniformity of Cell Shape (1-10)
5. Marginal Adhesion (1-10)

6. Single Epithelial Cell Size (1-10)
7. Bare Nuclei (1-10)
8. Bland Chromatin (1-10)
9. Normal Nucleoli (1-10)
10. Mitoses (1-10)
11. Class: (2 for benign, 4 for malignant)

Note that all the variables, apart from the diagnoses and the (unnecessary) ID, are in the same range (i.e. 1-10).

Let's add these names to the data set:

```
names(cancerData) <- c("id", "clumpThickness", "uniformityOfCellSize",
                       "uniformityOfCellShape", "marginalAdhesion", "singleEpithelialCellSize",
                       "bareNuclei", "blandChromatin", "normalNucleoli", "mitoses", "class")
```

Let's check the data to see if we have any issues

```
str(cancerData)
```

```
## 'data.frame':    698 obs. of  11 variables:
##  $ id                       : int  1002945 1015425 1016277 1017023 1017122 1018099 1018561 1033078 103
##  $ clumpThickness           : int  5 3 6 4 8 1 2 2 4 1 ...
##  $ uniformityOfCellSize     : int  4 1 8 1 10 1 1 1 2 1 ...
##  $ uniformityOfCellShape    : int  4 1 8 1 10 1 2 1 1 1 ...
##  $ marginalAdhesion         : int  5 1 1 3 8 1 1 1 1 1 ...
##  $ singleEpithelialCellSize: int  7 2 3 2 7 2 2 2 2 1 ...
##  $ bareNuclei               : chr  "10" "2" "4" "1" ...
##  $ blandChromatin           : int  3 3 3 3 9 3 3 1 2 3 ...
##  $ normalNucleoli           : int  2 1 7 1 7 1 1 1 1 1 ...
##  $ mitoses                  : int  1 1 1 1 1 1 1 5 1 1 ...
##  $ class                    : int  2 2 2 2 4 2 2 2 2 2 ...
```

It appears we do have some problems:

- The ID variable is irrelevant. We can probably remove it.
- The "bare nuclei data" has been interpreted aswhich suggests the presence of invalid values.
- The class(benign or malignan) is represented by 2 and 4, and we should probably recode that.

Let's clean up the dataset a bit. First, we remove the IDs:

```
cancerData$id <- NULL
```

Next, let's make the "bare nuclei" data numeric. Any values that can't be converted into the numeric data will be receive the values of 'NA', which R interprets as missing data it can ignore. We do so with the as.numeric function:

```
cancerData$bareNuclei <- as.numeric(cancerData$bareNuclei)
```

```
## Warning: NAs introduced by coercion
```

Because we will be running predictions, and we can't predict on missing data, we need to remove all rows with missing data. This is done relatively easily in R:

```
cancerData <- cancerData[complete.cases(cancerData), ]
```

Let's also change "class" values 2 and 4 into "for the data"clean-up" stage, let's transform classes of 2 and 4 into "benign" and "malignant" and turn the data into a factor:

```
cancerData$class <- factor(ifelse(cancerData$class == 2, "benign", "malignant"))
```

Let's check on the data again:

```
str(cancerData)
```

```
## 'data.frame':    682 obs. of  10 variables:
##  $ clumpThickness          : int  5 3 6 4 8 1 2 2 4 1 ...
##  $ uniformityOfCellSize    : int  4 1 8 1 10 1 1 1 2 1 ...
##  $ uniformityOfCellShape   : int  4 1 8 1 10 1 2 1 1 1 ...
##  $ marginalAdhesion        : int  5 1 1 3 8 1 1 1 1 1 ...
##  $ singleEpithelialCellSize: int  7 2 3 2 7 2 2 2 2 1 ...
##  $ bareNuclei              : num  10 2 4 1 10 10 1 1 1 1 ...
##  $ blandChromatin          : int  3 3 3 3 9 3 3 1 2 3 ...
##  $ normalNucleoli          : int  2 1 7 1 7 1 1 1 1 1 ...
##  $ mitoses                 : int  1 1 1 1 1 1 1 5 1 1 ...
##  $ class                   : Factor w/ 2 levels "benign","malignant": 1 1 1 1 2 1 1 1 1 1 ...
```

This looks much better. We now have 682 complete cases, which we need to separate into the training and the testing set. We can split it a number of ways, for example, 70/30 or 60/40. Let's split it into approximately 70/30.

First, split the predictor variables into training and test predictor sets:

```
trainingSet <- cancerData[1:477, 1:9]
testSet <- cancerData[478:682, 1:9]
```

Next, split the diagnoses (benign or malignant) into training and test outcome sets:

```
trainingOutcomes <- cancerData[1:477, 10]
testOutcomes <- cancerData[478:682, 10]
```

Assignment

To receive the grade of 70 points, you need to do the following:

1. Download and install the library called "class"
2. Familiarize yourself with the k-nearest neighbors classification
3. Perform the k-nearest neighbors classification
   - Hint: command is "knn"
   - Run it on the training set and training outcomes
   - Set the number of neighboring data points (the k) to be 21 (the square root of the number of training examples (477)
   - Command should be as follows:

     "'r library(class) predict <- knn(train = trainingSet, cl = trainingOutcomes, k = 21, test = testSet) "'
4. Familiarize yourself with the contents of the "predict" object
5. Evaluate the effectiveness of the model
6. Prepare the report on your findings

To receive the additional 30 points, you need to do the following:

1. Create a confusion matrix. How many cases were predicted correctly for the benign and the malignant types?
2. Use a different number for the k and re-run the model. Did you obtain a different result?

Prescriptive analytics assignment (Lecture 6)

This assignment is designed to help you get more acquianted with the linear programming models. You are welcome to use any software you would like for the calculations.

The assignment

You are given 3 linear programming problems.

To receive a grade of **70 points**, choose one problem.

To receive the additional **30 points**, choose any two.

For each of the problems, please do the following:

1. Formulate a linear programming problem (objective function, constraints, etc.)
2. Solve the linear program. It would be the easiest if you use the code from Lecture 6, but feel free to solve it any way you prefer
3. Submit your problem formulations and your solutions in one document

The choice of problems

Here are the problems you can choose from.

1. This is a profit maximization problem.
    - A product can be made in three sizes: large, medium, and small, which yield a net unit profit of \$15, 12, and 10, respectively.
    - The company has three centers where this product can be manufactured and these centers have a capacity of turning out 450, 950, and 375 units of the product per day, respectively, regardless of the size or combination of sizes involved.
    - Manufacturing this product requires cooling water and each unit of large, medium, and small sizes produced require 23, 19, and 7 gallons of water, respectively.
    - The centers 1, 2, and 3 have 12,000, 6000, and 3500 gallons of cooling water available per day, respectively.
    - Market studies indicate that there is a market for 900, 700, and 250 units of the large, medium, and small sizes, respectively, per day.
    - By company policy, the fraction (scheduled production)/(center's capacity) must be the same at all the centers.
2. This is a cost minimization problem
    - A farmer is raising pigs for market. He needs to determine the quantities of the available types of feed that should be given to each pig to meet certain nutritional requirements at a minimum cost.
    - The number of units of each type of basic nutritional ingredient contained in a kilogram of each feed is given in the file "LP3.csv" in your project folder.
    - Formulate and solve the linear programming model for this problem
3. This is a profit maximization problem.

- An electronics company has a contract to deliver 26,350 radios within the next four weeks.
- The client will pay:
  - $25 for each radio delivered by the end of the first week
  - $20 for those delivered by the end of the second week
  - $18 by the end of the third week
  - $12 by the end of the fourth week.
- Each worker can assemble only 50 radios per week
- The company only employs 45 people
- To meet the order, it must hire and train temporary help. Any of the experienced workers can be taken off the assembly line to instruct a class of three trainees; after one week of instruction, each of the trainees can either proceed to the assembly line or instruct additional new classes.
- Currently, the company has no other contracts, so some workers may become idle once the delivery is completed.
- All of them, whether permanent or temporary, must be kept on the payroll until the end of the fourth week.
- The weekly wages of a worker, whether assembling, instructing, or being idle, are $220
- The weekly wages of a trainee are $110.
- The production costs, excluding the worker's wages, are $5 per radio.
- Formulate this as an LP problem

Thank you

I would like to thank you for staying with me throughout this course. This course is very unusual, both in format and in content. I am glad that you have completed it despite your busy schedule; I am sure it will prove very useful to your career. Hope to see you in our future courses or in our Master's program!

Again, thank you for taking my course.