Statistics & Methodology
Group Project

SUBMISSION DEADLINE: 11 October 2019 at 23:59

1 Introduction

You will analyze data from Wave 6 of the *World Values Survey*. These data are freely available online, but you must access them yourself (I cannot re-distribute the data).

1.1 Data Access

- To access the data, follow these steps:
  1. Follow this link:
     `http://www.worldvaluessurvey.org/WVSDocumentationWV6.jsp`
  2. Download the file named: **WV6_Data_R_v20180912** from the "Statistical Data Files" cell in the table.
     - You will be re-directed to a page from which you can initiate the download.
  3. Under "PERSONAL DATA," provide the requested information.
  4. Under "FILE USAGE," give an appropriate project title and description (just explain that you'll use the data for a class project), and set the "Intended use" field to *Instructional.*
  5. Check the box labeled: *I have read the 'Conditions of use' and agree with them*, and hit the *Download* button.
     - The downloaded file should be a ZIP archive.
  6. Extract the contents of the downloaded ZIP archive into the *data* subdirectory of the directory tree for this project.
     - You should now have the data saved as an RDS file in the *data* subdirectory.

1.2 Data Processing

- Once you have downloaded the data, you must process them by running the **processWvsData.R** script located in the *code* subdirectory of the directory tree for this project.
  - For the `dataDir` argument, you will need to specify the relative file path to the *data* directory.

  - For the `fileName` argument, you will need to give the filename of the downloaded RDS file (don't forget the file extension).
  - After defining these two variables, execute the entire script.
  - The processed data will be saved in your *data* subdirectory as **wvs_data.rds**.
- You will use these processed data for all of the analyses requested in Section 2.

1.3 Additional Information

- The data do not have informative variable names (either before or after processing). You will need a codebook to decipher the variables' meanings.
  - The appropriate codebook is located in the *docs* subdirectory of the directory tree for this project.
  - The codebook file is named: **WV6_Codebook_v20180912.pdf**.
- Unless otherwise noted, assume an α-level of $\alpha = 0.05$ for all significance tests.
- Unless otherwise noted, all prediction errors should be quantified in terms of *mean squared error* (MSE).
- For the purposes of this project, you may treat Likert-type items as continuous variables.
- In the following section, the number of points possible for each question is given in brackets after the question.

2 Questions

This section contains the questions you will answer for this assignment.

2.1 Data Cleaning

The questions in Sections 2.3 and 2.4 will require you to select a number of variables for your analyses. In this section, you will clean those variables.

1. List the variables you will use for the analyses in sections 2.3 and 2.4. **[2]**
   - For each item provide the following:
     (a) The variable's name on the dataset (e.g., V4)
     (b) A human-readable description of the variable (e.g., The importance of family)
     (c) A list the questions for which you will use that variable

NOTE: Unless otherwise specified, all remaining questions in this document pertain to only the variables you listed in Question 1.

2. What is the proportion of missing data for each variable? **[1]**
3. Check each variable for *univariate* outliers. **[3]**
   (a) In one or two sentences, justify your choice of outlier test (i.e., describe why you used your chosen method).

   (b) For each variable, report the row indices of any outliers you detected.
4. Treat the outliers you detected in Question 3 in some reasonable way. **[2]**
   (a) What method did you use to treat the outliers?

   (b) In one or two sentences, justify the treatment method you applied (i.e., describe why you addressed the outliers in the way you did).
5. Check each row of the dataset for *multivariate* outliers. **[3]**
   (a) In one or two sentences, justify your choice of outlier test (i.e., describe why you used your chosen method).

   (b) Report the row indices of any outliers you detected.

   (c) Exclude outlying observations from the data.

Note: Unless otherwise specified, all following analyses should be conducted on the cleaned data you produced in this section.

2.2 Exploratory Data Analysis

1. Which countries are represented in these data? **[1]**
2. What are the sample sizes for each country represented in these data? **[1]**
3. Report the means, medians, standard deviations, and ranges of each continuous variable. **[2]**
4. Report the frequency table of each categorical variable. **[1]**
5. Generate histograms for each variable. **[2]**
6. Generate kernel density plots for each continuous variable. **[3]**
7. Based on the results of this EDA, did you notice anything about these variables that may impact the analyses you will do in Sections 2.3 and 2.4? **[2]**
   - Briefly describe what you noticed (if anything) and the potential implications for your inferential/predictive analyses.

2.3 Multiple Linear Regression

1. Is there a significant effect of country on feelings of happiness (*happiness*)? **[2]**
2. Which country is the <u>most</u> happy? **[1]**

3. Which country is the least happy? **[1]**
4. Is there a significant effect of subjective state of health (*health*) on *happiness* after controlling for country? **[2]**
5. Which country is the most happy, after controlling for *health*? **[1]**
6. Which country is the least happy, after controlling for *health*? **[1]**
7. Briefly describe (i.e., in one or two sentences) how the country-specific levels of *happiness* change after controlling for *health*. **[2]**
    - The models that do and do not control for *health* will imply somewhat different patterns of *happiness* effects. Give a qualitative description of how the *happiness* effects change when controlling for *health*.

2.4 Predictive Modeling

In this section, you will be building linear regression models to predict peoples' reported satisfaction with the financial situation of their household ($FinSat$).

1. Select and list three plausible, non-nested sets of predictors (or functions thereof, e.g., interactions or polynomials) to use in predicting $FinSat$. **[2]**
    - You do not need to justify your selection with literature references/prior research (although you are free to do so, if you want).
    - I really just want you to use common sense/intuition to select three different groups of variables that could be expected to predict $FinSat$.
    - Think about this task from a model comparison perspective. Try to come up with three sets of variables that reflect three unique predictive processes that may be interesting to compare.
2. Briefly explain your rationale for choosing the three sets of predictors you described in Question 1. **[2]**
3. Use 10-fold cross-validation to compare the predictive performance of the three models define in Question 1. **[4]**
    (a) Report the cross-validation error (CVE) from each model.
4. Which of the three models compared in Question 3 performed best? **[2]**
5. What is the estimated prediction error of the model you selected in Question 4? **[2]**

3 Write-Up

This section describes the documentation you must submit for this assignment.

3.1 Written Report

You will submit a single document containing (clearly numbered) answers to each of the questions in Section 2.

- This document must be in PDF format.
- Each answer should be brief. One or two sentences will suffice in most cases.
- Waffling will not help you. If part of your answer is correct, but another part contradicts and/or invalidates the correct part, you will not receive points for the correct information.

Where appropriate, answers must be supported with in-text statistical information (as you would see in an scientific journal article). Consider the following example:

Question: Is there a significant effect of age on BMI?

Answer 1: Yes

Answer 2: No

Answer 3: Yes ($\beta = 0.5$, $SE = 0.125$, $t = 3.33$, $p < 0.001$)

Answer 4: No ($\beta = 0.5$, $SE = 0.35$, $t = 1.43$, $p = 0.156$)

In this case, Answers 1 and 2 would receive no points whereas Answers 3 and 4 would receive full credit (assuming they were otherwise correct).

- Note that a p-value alone is not sufficient statistical justification.
  - When answering question that ask if a given effect is statistically significant, you must report, at least, the following:
    1. The estimated effect (e.g., $\hat{\beta}$, R^2, ΔR^2)
    2. The test statistic (e.g., t, F) or confidence interval (CI)
    3. The p-value (if not using a CI)
  - When reporting test statistics and p-values, the standard error should also be given, if it is provided by the software.
- Some questions clearly do not require (or admit) statistical justification (e.g., Question 1 from the *Predictive Modeling* section). In these situations, you should not include any statistical results in your answer.

3.2 Analysis Syntax

As is true of all professional data analyses, you must submit a complete script that executes all of the analyses you used to complete this assignment.

- This syntax will contribute **20 Points** to your assignment score.
- All syntax files must be plain text files with a *.R* file extension.

- Do not embed code snippets directly in your written report.
- To receive full credit, your script must satisfy the following conditions:
    1. It must run, without errors, in "batch-mode" (i.e., without any manual input, editing, or modification from the user).
    2. It must produce each result necessary to answer the questions asked in Section 2.
- Failure to fully satisfy either of the two preceding conditions will results in lost points.

To receive credit for the answers in your written report, the results returned by your code must match the results reported in your write-up, after allowing for rounding errors.

- Your syntax should be annotated to clearly indicate which section of code corresponds to which questions.
- Answers that cannot be directly linked to executable code (that provides output supporting the written answer) will receive no credit.
    - The obvious exceptions to this rule are questions that can be answered without any type of programming or data analysis.

3.2.1 Practical Tips

- Make sure to convert categorical variables into factors before using them in your analysis.
- Make sure to set the random number seed. If you don't set the seed, any code that uses random number generation (e.g., cross-validation) will produce different results each time it is run.
- Execute your entire syntax file from beginning to end (i.e., source the entire script) to produce the results included in your written report.
    - This is how I will run your code while grading.
    - If you rerun the same code chunk multiple times, the state of the random number seed will advance beyond the version I will be using for grading.
- You may assume that I have all necessary packages installed on my machine; your syntax does not need to include code to install the packages you use in your analysis.
    - You do, however, need to use the `library()` function to load any packages that you will use in your analysis.

4 Submission Procedure

Each group will upload a single submission via Canvas.

- The group project grade will apply equally to each member of the group.

- There will be no weighting based on the relative contribution of the group members.

You will submit your project as a single ZIP archive.

- This ZIP archive will be named: *groupN.zip* (where *N* is replaced by your group number).
  - E.g., the ZIP archive for Group 42's submission will be *group42.zip*.
- This ZIP archive will containing the following:
  - A single parent directory called *groupN* (where *N* is replaced by your group number).
    - E.g., the parent directory for Group 42's submission will be *group42*.
  - The *groupN* parent directory will contain the following subdirectories:
    - A *code* directory containing all of your syntax files
    - A *data* directory containing all of your data files
    - A *docs* directory containing your written report and any supporting documentation that you wish to provide
    - A *figs* directory containing any graphics that you create (these should also be embedded in your written report)

The written report must be provided in PDF format.

- The names, student numbers (SNRs), and administrative numbers (ANRs) of each group member must appear on the first page.

All syntax files must be provided as plain text with a *.R* file extension.

- The names, SNRs, and ANRs of each group member must appear on the first page of each file.

5 Grading

This section explicates the procedures used to compute your grade on the assignment.

5.1 Grading Scheme

This assignment will be worth a maximum of **70 Points** with the following distribution:

- Written Report: 45 Points
- Analysis Syntax: 20 Points
- Formatting: 5 Points

Your final grade on this assignment will be determined by summing the number of points scored, dividing this total by 7, and rounding the result to two decimal places.

5.2 Grading Procedure

To evaluate your project, I will unzip the *groupN.zip* archive on my computer and run your code without modifying any syntax or moving any files/directories.

- To receive full credit, all of your analyses must run in this self-contained fashion.

Assume that the working directory will be the *code* subdirectory of your submission. This assumption will hold when I run your syntax for grading, so anything you do to override this assumption (e.g., manually setting the working directory) will likely lead to lost points.

- **DO NOT** manually set your working directory with the `setwd()` function (or via any other means).
- **DO** use relative paths anchored in the *code* subdirectory to read in data files and write out figures.

As noted in the rubric below, the correctness of your answers will be evaluated with respect to the results that I get when running your analysis syntax. So, ensuring that your code runs in a portable fashion (i.e., not just on the machine where it was written) is a necessary condition for receiving full credit in this assignment.

5.3 Rubric

The written report will be graded as follows:

- Correct answers that can be linked to your analysis syntax will get full credit.
- Incorrect answers that can be linked to your analysis syntax will get partial credit.
- Correct questions that cannot be linked to the analysis syntax will get no credit.
- If your analysis syntax provides two conflicting answers to the same question, you will get no credit for that question (even if one of the conflicting answers is correct).
- Correct answers that do not provide sufficient statistical justification (when appropriate) will receive no credit.
- Answers that contain the correct answer as well as contradictory incorrect information will be treated as incorrect.
  - If part of your answer is correct but a different part contradicts the correct part, your answer will be regarded as incorrect.

The analysis syntax will be graded as follows:

- Everyone begins with the full 20 points.
- Five (5) points will be deducted if the data cannot be loaded without modifying the syntax.
- Two (2) points will be deducted for each object that is used by the code without being initialized (which will throw an error).

    - E.g., if your code tries to do something to `dat1` before `dat1` is created, you will lose two points.

- One (1) point will be deducted for each other type of error (i.e., error message, not incorrect analysis) returned by your code.
- One (1) point will be deducted for each extraneous chunk of code (i.e., pieces of code that clearly have no bearing on the group project analysis).

The five (5) possible points for formatting are distributed as follows:

- Two (2) points for submitting your project as a correctly named ZIP archive with the correct directory structure
- One (1) point each for:
    - Using the correct file types for your written report and analysis syntax
    - Including all group member names on the written report and analysis syntax files
    - Including all group member SNRs and ANRs on the written report and analysis syntax files