Statistics & Methodology
Group Project

SUBMISSION DEADLINE: 10 October 2018 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_v_2016_01_01** from the "Older version of Data Files" cell in the table.
     - You will be re-directed to a page from which you can initiate the download.
     - Note that this is not the most recent version of the data file. Make sure to download the correct file.
  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.
  6. Save the downloaded RData (i.e., R Workspace) file in the *data* subdirectory of the directory tree for this project.

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 also need to give the filename of the downloaded RData file.
  - 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 below.

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: **F00007761-WVS6_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).
- In the following section, the number of points possible for each question are given in brackets after the question.

2 Questions

2.1 Multiple linear regression

1. Which countries are represented in these data? **[1]**
2. What are the sample sizes for each country represented in these data? **[1]**
3. Overall, is there a significant effect of country on feelings of happiness? **[2]**
4. Which country has the highest level of feelings of happiness? **[1]**
5. Which country has the lowest level of feelings of happiness? **[1]**
6. How do the country-specific levels of feelings of happiness change after controlling for subjective state of health? **[2]**

2.2 Continuous variable moderation

1. After controlling for country, does the importance people afforded to democracy (*DemImp*) significantly predict the extent to which they think their country is being run democratically (*DemRun*)? **[2]**

2. After controlling for country, does the $DemImp \rightarrow DemRun$ effect vary as a function of peoples' satisfaction with their lives (SWL)? **[2]**
3. Within what range of SWL is the $DemImp \rightarrow DemRun$ simple slope from Question 2 statistically significant? **[3]**

2.3 Categorical variable moderation

1. After controlling for SWL, does the $DemImp \rightarrow DemRun$ effect vary significantly by country? **[2]**
2. Visualize the results from Question 1 in a suitable way. **[3]**
3. For which country is the effect of $DemImp$ on $DemRun$ strongest, after controlling for SWL? **[3]**
4. For which country is the effect of $DemImp$ on $DemRun$ weakest, after controlling for SWL? **[3]**
5. Are the simple slopes referenced in Questions 3 and 4 statistically significant? **[3]**

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 (theoretically justified) sets of predictors (or functions thereof, e.g., interactions or polynomials) to use in predicting $FinSat$. **[1]**
2. Briefly explain why you expect the three sets of predictors you chose in Question 1 to perform well. That is, explain your rationale for defining these three sets. **[2]**
3. Use 10-fold cross-validation to compare the predictive performance of the three models define in Question 1. **[4]**
4. Which of the three models compared in Question 3 performed best? **[2]**
5. What is the estimated prediction error of the best model? **[2]**
6. Based on the selection you made in Question 4, what can you say about the attributes that are are important for predicting financial satisfaction? **[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 preceding questions.

- 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).

- 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 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 every question asked above.
- 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 so as to clarify which section of code corresponds to which questions.
- Answers that cannot be directly linked to executable code (that provides output matching 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.

4 Grading

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

4.1 Grading Scheme

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

- Written Report: 42 Points
- Analysis Syntax: 20 Points
- Formatting: 8 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 Submission Procedure

Each group will submit a single project.

- The grade for that project 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 *.zip* archive with the following structure:

- A single parent directory that contains:
  - A *code* subdirectory containing all of your syntax files
  - A *data* subdirectory containing all of your data files
  - A *documents* subdirectory containing your written report and any supporting documentation that you wish to provide
  - A *figures* subdirectory containing any graphics that you create (these should also be embedded in your written report)

To evaluate your project, I will unzip this 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.

The written report must be provided in *.pdf* format.

- The names, student numbers, and administrative numbers 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, student numbers, and administrative numbers of each group member must appear on the first page of each file.

To receive all of the 8 points possible for “formatting,” your submission must fully satisfy each of the conditions noted in this section.