HAD5772: Final Project Guidelines

# Data Analysis Project and Rubric

## Overview

This project gives you hands-on experience summarizing and analyzing data of your own interest using R. You should use publicly available healthcare data to address a research question of interest, as discussed in class. Feel free to consult with me or with other faculty if you need help finding a specific type of data.

This project prepares you to answer questions such as:

1. Measurement: On average, how big is variable ?
2. Heterogeneity: How widely does vary?
3. Correlation (causality?): Are variables and positively or negatively correlated? How strong is this relationship?
4. Prediction: How can I use variable to predict variable ?
5. Recalibration: Which observations of are larger or smaller than would be predicted using ?

## Part 1 – Data Collection & Summary (+58)

You may submit your answers to the questions below simply as a numbered list.

1. (+19) *Basic data description*
   1. (+15) Collect data of interest, and state where you got it. You do *not* need to submit your data files.
   2. (+1) How were these data originally collected, and by whom?
   3. (+1) What is your unit of observation? How many observations do you have? [[1]](#footnote-1)
   4. (+2) List at least one variable that is *quantitative* (e.g. price, number of sales, age, GDP) and another that is *binary* (e.g. gender, race, industry, political party, sport position).[[2]](#footnote-2) If any of these variables are not obvious already, explain how they are determined or measured (e.g. what units) [[3]](#footnote-3) or constructed.

Note that, while not required, it is often interesting to pull data from multiple sources, or to construct new variables from existing data. In the spreadsheet below, for example, government finance variables come from one source and a binary political variable comes from another. Per capita variables are then computed simply as ratios; growth variables are computed simply as differences (as a ratio of the original level); and additional binary variables are constructed either by reducing a quantitative variable into “high” and “low” categories (e.g. GDP growth above or below 1.5%) or by comparing two existing variables (e.g. Gov. growth > GDP growth?).

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Unit | Original Variables | | | | Constructed Variables | | | | | |
|  | GDP | Population | Gov. Spending | Republican House? | Per capita GDP | Per capita GDP growth | GDP Growth > 1.5%? | Per capita Gov. spending | Per capita Gov. growth | Gov. growth > GDP growth? |
|  |
|  |
|  | ($ bil.) | (mil.) | ($ bil.) |  | ($ thous.) | (%) |  | ($ thous.) | (%) |  |
| Year |  |  |  |  |  |  |  |  |  |  |
| 2008 | 14,834 | 304 | 4,665 | 0 | 48.8 | - | - | 15.3 | - | - |
| 2009 | 14,418 | 307 | 5,179 | 0 | 47.0 | -3.7% | 0 | 16.9 | 10.1% | 1 |
| 2010 | 14,779 | 309 | 5,057 | 0 | 47.8 | 1.7% | 1 | 16.3 | -3.2% | 0 |
| 2011 | 15,052 | 312 | 5,116 | 1 | 48.3 | 1.1% | 0 | 16.4 | 0.4% | 0 |
| 2012 | 15,471 | 314 | 5,042 | 1 | 49.3 | 2.0% | 1 | 16.1 | -2.2% | 0 |
| 2013 | 15,759 | 316 | 4,955 | 1 | 49.8 | 1.1% | 0 | 15.7 | -2.5% | 0 |
| 2014 | 16,077 | 319 | 4,957 | 1 | 50.4 | 1.3% | 0 | 15.5 | -0.7% | 0 |

1. (+3) *Identify your* a*udience*

Identify some audience that might find these data interesting: a policy maker, a business or industry leader, a consumer, etc. In Part B, you will report your findings to this audience. List any questions (at least two) that this audience might have, that you believe your data can shed (at least partial) light on.

1. (+9) *Summarize individual variables*
   1. (+1) Summarize at least one binary variable by reporting the total fraction of observations in each category.
   2. (+3) Summarize at least one quantitative variable by reporting the minimum and maximum, mean, median, and standard deviation.
   3. (+2) Use one binary variable to divide your data into subgroups, and report the conditional minimum, conditional maximum, conditional mean, and conditional standard deviation for this subgroup (e.g. average wages among female workers). Note: for all subsequent analysis of this project, you may use the full sample, or this restricted sample, as you wish.
   4. (+3) Represent at least one quantitative variable graphically, using a histogram.
2. (+10) *Correlation and causation*

Choose two variables, and do the following:

* 1. (+8) Identify reasons why the variables might be positively or negatively correlated. Might one cause the other to increase or decrease? Is reverse causation possible? Are there outside factors that might cause both variables to move? If so, what additional data could be collected and examined, to control for these outside factors?
  2. (+1) Based on the possible causation described above, make a guess of what the sign and magnitude of the correlation coefficient between these variables will be.
  3. (+1) Compute the actual correlation coefficient and compare it with your prediction above.

1. (+6) *Graphical Summary*
   1. (+2) Make a histogram of one of your quantitative variables. Briefly describe some facet of the distribution that is evident in the histogram.
   2. (+4) Compare two variables graphically, using something like the following. Include labels (e.g. color-code, axis labels, legend, etc.) and ensure that your figure is well-formatted. Briefly describe some facet of the relationship between the two variables that is apparent in the type of graphic you chose.
2. (+11) *Regression*

Regress one variable on another variable , and answer the following.

* 1. (+3) Report the slope and intercept parameters for your regression. Interpret your slope parameter in the context of your two variables.
  2. (+2) State and interpret the associated with your regression.
  3. (+2) Which observation of is the furthest above the regression line? Which is the furthest below?
  4. (+2) Report the residual standard deviation .
  5. (+2) Predict for some special value of of your choice.

## Part 2 – Executive Summary (+29)

Write a short[[4]](#footnote-4) report or memo, addressed to the audience identified above. Summarize and explain any patterns in your data that you find interesting or useful. You may, but are not required to, include the statistics or graphs that you produced above, or additional statistics or graphics that you have derived. In your report, you should do the following:

1. (+2) Clearly state the question or issue that this analysis addresses.
2. (+2) Make sure the nature of the data, including key variables, is clear to the executive.
3. (+6) Clearly explain key findings. (Graphical representations may be helpful here.)
4. (+3) Explain and emphasize the practical significance of any key results. Include any policy recommendations (e.g. shopping strategies, legal regulations, etc.) that your analysis favors.
5. (+6) Be clear and forthright about any caveats, assumptions, or limitations of your data, your analysis, or your policy recommendations, including questions of causation. Indicate what additional data or analysis would be necessary to answer the questions of interest more completely.
6. (+5) Organize your thoughts effectively.
7. (+5) Write ideas clearly, cleanly (e.g. without grammatical errors), directly, and succinctly (executive audiences don’t have time for tangents and roundabout arguments). As if your audience has only a limited knowledge of statistics, avoid overly technical jargon. (For example, units of dollars are easier to understand than standard deviations or correlation coefficients.)

## Part 3 – Statistical Inference (+26)

Do the following, stating any important assumptions that your answers rely on.[[5]](#footnote-5) You do not need to write out all your computations, but should make clear how you arrived at your answers.

1. *Mean*
   1. (+2) For at least one quantitative variable, compute a confidence interval for the underlying population mean at a confidence level of your choice.[[6]](#footnote-6)
   2. (+2) Perform a one- or two-sided test, at the level of your choice, of the hypothesis that is equal to a benchmark value of your choice. State the associated p-value.
2. *Standard Deviation*
   1. (+2) For at least one quantitative variable, compute a confidence interval for the underlying population standard deviation , at a confidence level of your choice.[[7]](#footnote-7)
   2. (+2) Perform a one- or two-sided test, at the level of your choice, of the hypothesis that is equal to a benchmark value of your choice. State the associated p-value.
3. *Proportion*
   1. (+2) For at least one binary variable, find a point estimate of the underlying proportion in a particular category. Compute a confidence interval for, at a confidence level of your choice.[[8]](#footnote-8)
   2. (+2) Perform a one- or two-sided test, at the level of your choice, of the hypothesis that is equal to a benchmark value of your choice. State the associated p-value.
4. *Difference of Means*
   1. (+2) Use a binary variable to divide your data into two subgroups. For at least one quantitative variable, find a point estimate of the difference between the means of the underlying subpopulations. Compute a confidence interval for, at a confidence level of your choice.
   2. (+2) Perform a one- or two-sided test, at the level of your choice, of the hypothesis that equals a benchmark value of your choice. State the associated p-value.
5. *Regression*
   1. (+3) Regress one variable on another variable . Report point estimates and of the intercept and slope coefficients and the coefficient of determination .
   2. (+2) Identify which two observations are the furthest above and furthest below the regression line.
   3. (+1) Report the estimated standard deviation associated with the error term .
   4. (+2) Give a confidence interval for , at the level of your choice.
   5. (+2) Perform a one- or two-sided test, at the level of your choice, of the hypothesis that equals a benchmark value of your choice. State the associated p-value.

1. You need at least three observations; larger samples increase precision. If you have trouble identifying the unit of observation, it may be that your data summarize more primitive raw data. The number of observations may effectively be one in that case, making your data unusable. [↑](#footnote-ref-1)
2. When the unit of observation is a time stamp (e.g. day, year, or week), it can also double as a quantitative variable. You can make a categorical variable binary simply by combining categories. For example, a “race” variable might have several codes for different races but can be reduced simply to “white” and “non-white”. You can also construct binary variables from quantitative variables (see below). [↑](#footnote-ref-2)
3. For example, a humanitarian agency might rate sovereign governments as “corrupt” or not, and designate individuals as “in poverty” or not, but how are these categories assigned? What exactly do they mean? [↑](#footnote-ref-3)
4. There is no required length. Your goal is to be as clear, informative, and concise as possible. [↑](#footnote-ref-4)
5. This section presumes that your data represent a sample from a larger population of interest. If your data represent an entire population (e.g. all 50 of the United States), merely perform the following analysis *as if* it were only a sample from a much larger population (e.g. a huge pool of U.S. states, from which you have randomly drawn 50). [↑](#footnote-ref-5)
6. If you wish to type equations or mathematical symbols in Word, hold the “alt” key and type the sign to open the equation editor. For Greek variables, type and then spell the name of the letter (e.g. \mu for , \sigma for ). can be typed as x\bar. The T.DIST and T.INV commands in Excel and “tprob” and “invt” commands in Stata give you cdf values and inverse cdf values for a t-distribution. Stata’s “ci” confidence interval function also has an option for means. [↑](#footnote-ref-6)
7. CHISQ.DIST and CHISQ.INV commands in Excel or “chi2” and “invchi2” commands in Stata give Chi-square cdf and inverse cdf values. The “ci” function in Stata has an option for variances. [↑](#footnote-ref-7)
8. The NORM.DIST and NORM.INV functions in Excel and the “normal” and “invnormal” functions in Stata give Normal cdf and inverse cdf values. [↑](#footnote-ref-8)