SMI606 Introduction to Quantitative Research: Assessment 2

Dr Calum Webb

2023-08-25

Assessment outline

For this assessment you must submit a 2,000 word research paper on a topic of your choice which uses at least **one** of the following quantitative methods in R: (a) multiple linear regression; (b) multiple logistic regression; (c) cluster analysis; (d) spatial analysis. It is up to you whether you wish to use one of the datasets provided on Blackboard, or whether you would like to use some other secondary data you have found. **You must not collect primary data (data collected on your own, e.g. conduct a survey, do an observation, etc.) without ethical approval**.

This research paper should include an introduction and rationale; a research question (or questions); a brief literature review; a brief overview of the data and method(s) used; a section detailing the findings, including appropriate output from, and interpretation of, the method(s) of choice; a brief discussion section which also includes any limitations or checking of assumptions; and a conclusion.

The expectation is that you only use one of the above methods, *however*, you may use more than one if you think it would be beneficial to the content of your paper (for example, if you chose to do spatial analysis you might also want to include a multiple regression model to talk about some covariates). You will not be penalised or rewarded for using more than one method, but trying to do too much may mean that you struggle to cover everything in sufficient detail.

More details are provided about each of these steps in the second part of this document ("Detailed Assessment Guidance").

Word Limit

2,000 Words

Tables and graphs are not included in the word limit, though they should be used appropriately (there must be a good reason if long sections of text are included in a table). There are no limits on the number of tables or graphs you can use, but the content of all graphs and tables should be discussed and pertinent parts should be at least briefly described in the main text.

R code, comments, and output is also not included in the word limit, however, all pertinent information should always be in the main body of your assigment. You can use the Rstudio `wordcountaddin` Add In to get an accurate word count of your assignment if it is written in Rmarkdown, which you are encouraged to use (requires installation with devtools: devtools::install_github("https://github.com/benmarwick/wordcountaddin")). Comments should be restricted to explaining the purpose of the code, and not used for communicating or interpreting the results. Bibliographies are not included in word counts, but in-text citations are.

Assessment Value

70% of final grade for SMI606

Deadline

See Turnitin item on Blackboard (SMI606 -> Assessment menu)

References and formatting

You may choose whichever referencing format you wish (e.g., APA, Chicago, etc.), but please be consistent throughout your submission.

Completed assessments should be submitted via Turnitin as a .pdf or a .docx file. You are strongly encouraged to complete your assessment in Rmarkdown and submit the .pdf output. You can use the template .Rmd file provided on the assessments page on Blackboard.

You should include your `R` code as well as the output, ideally just before each output — this is why completing the assessment in Rmarkdown can be much easier. For example, if you report a correlation in your writing you should have the code used to calculate the correlation and the result from that code somewhere above the paragraph where it is mentioned. `R` code will be spot-checked to ensure that it matches up with the output.

Submission

Coursework must be submitted online through Turnitin and by no later than 12:00pm (noon) on the day of the deadline. Any unauthorised late submissions after midday on the day of the deadline will incur a penalty of up to 100%, as per the student handbook. Marked coursework will generally be returned within 3 working weeks. The pass mark for this module is 50% overall. Any change to assessment arrangements will be announced in Blackboard.

Assessments must not be submitted via email. They must only be submitted via Turnitin. Any other method of submission will not be marked.

Check the module outline and student handbook for further details about assessment submission and feedback.

Detailed Assessment Guidance: Writing your research paper

Below is further guidance for what is meant by each of the aspects expected to be included in the research paper. While this paper is 2,000 words in total, you will still find this limit reasonably constraining. As such, it's important for you to plan out how you wish to structure your report. I would recommend that you try to ensure you have at least 1,000 words for your findings, discussion, and conclusions.

Paper Structure: You do not have to structure your research paper in exactly the way I have broken it down. You may choose to combine some parts or further break up others and this may improve readability. It will depend on what you have chosen to do.

I would strongly recommend looking at examples of short form research articles as a way to think about how you would structure and write your research paper (e.g. see Webb, Bywaters, Elliott, & Scourfield, 2021 example on Blackboard, which is 3,000 words).

Tone: You are expected to write this report as a quality piece of academic work, as if you were submitting it to a journal or as if you were publishing it. It should not be informal or read like a draft or set of notes.

Code: You do not need to explain your code, and you certainly should not spend parts of your word count explaining what your code does and why it works. You may add comments to your `R` chunks in `Rmarkdown` to this effect, but you will not receive extra marks for this. The purpose of this piece of work is to test whether you can correctly run and interpret the statistical tests and methods covered in the module, not to test how well you understand `R`.

In this regard, it also makes no difference how well or poorly your code is formatted — you will not be penalised, even if your `R` code formatting makes me want to cry. As long as the code you write leads to the appropriate output, and as long as this output is correctly interpreted in the context of answering a social science research question, you will receive full marks.

Introduction, rationale and research question

Objective: Describe the research question being explored and provide a rationale for why you are exploring this research question.

The research question should relate to the variables you are using the research paper and the data you have chosen. It can be relatively broad but should be able to lead to a clear answer. A rationale is a justification for *why* you have chosen this research question. It is conventionally rooted in the academic literature or in a social or societal problem or policy domain. The rationale should establish why it is important to answer this research question and why there is currently a gap in the existing research evidence ('literature'). The research question does not have to come before the research rationale; traditionally, the rationale comes first as an introduction.

Literature Review

Objective: Critically reflect on existing evidence and where your research paper makes a novel contribution.

A good literature review is able to give the reader a clear sense of what has already been established in the research field and why the proposed research question is of interest. It often recounts similar studies which may have complementary or contradictory findings and describes some of their differences (e.g. their research design; the country they were conducted in; how something was measured; or the unit of analysis used — e.g. individuals as opposed to small areas). However, it may also draw on untested theories or logical progressions of such theories.

Even if there has been little or no research conducted on your chosen topic, there is almost definitely related literature that could be discussed. For instance, if you are researching something about social inequalities in

access to a public service that has never been explored before, you could still talk about the existing literature about social inequalities in health or some other outcome which may be a relevant way of describing why your topic also matters.

This literature review does not need to be very long at all (likely around 300-400 words maximum). It should demonstrate to the reader that you are able to place your own research into a wider academic context.

Description of Data

Objective: Describe the data being used.

You should always provide an adequate description of the data that you are using. This includes: its source; the number of observations included (these might be people, neighbourhoods, countries, or other things); whether it is a sample or complete population; how the sample was derived (whether it is random or not); and contextual information about the population it is derived from. In order to do this adequately in this assessment, you may need to explore the 'Source and Info' links provided in the description of datasets description file. It is often considered good practice to provide a table of summary statistics for the variables used in your paper.

Description of Method(s)

Objective: Provide a basic description of the method used for a general audience.

You should also include a very brief description of the method that you have chosen to use, which traditionally accompanies the description of your data. The purpose of this is **not** to explain how the method works for a reader, but to explain what it is called (e.g. "A multiple linear regression model was estimated to..."; "A multiple logistic regression model was estimated to...") and why it was chosen (e.g. why does it help you meet your outcome or aim?).

You should also describe how variables were measured in this section and any purpose for inclusion. For example, "This study used the rate of hospitalisations for drug misuse per 10,000 people in a small area as the dependent variable. The proportion of people aged 16-19 in the area, the median income, and the distance to the nearest hospital were used as independent variables. Distance to the nearest hospital was included in order to control for higher rates of drug misuse associated with proximity to services, rather than higher incidences of misuse alone". You should include details of any transformations or recoding of your variables here.

Findings

Objective: Present and report the research findings arising from your use of the chosen statistical method.

Remember, **you are only expected to use one of these methods, but you have the option to use whichever you wish**.

Multiple Linear Regression

The output from your multiple linear regression model should be presented in a table (or as summary output) which includes estimates for all included independent variables (plus intercept); standard errors; t-statistics; p-values; and R-squared value.

You should explain what the independent variables' estimated linear associations with the dependent variable mean. If you are using a large number of additional variables that are not necessarily of substantive interest,

but are things you feel like the model should 'control' for (e.g. potential confounders), you do not have to report these.

Do not simply report something like "The slope of the linear relationship between variable X1 and Y was 0.345"; explain what this means, e.g. "There was a positive association between variable X1 and Y where an increase of 1 in X1 was associated with an increase in Y of 0.345; this indicates that small areas with higher X1 also have higher incidences of Y".

You may wish to include partial plots using tools like `ggeffects` (e.g. https://strengejacke.github.io/ggeffects/articles/introduction_partial_residuals.html), but this is not an expectation. They can, however, be useful if you have transformed your dependent variable in some way that makes the coefficients difficult to describe (e.g. using the natural log).

Multiple Logistic Regression

The output from your multiple logistic regression model should be presented in a table (or as summary output) which includes estimates for all included independent variables (plus intercept); standard errors; test statistics; p-values; and R-squared value. You should make it clear whether you are including change in logged odds, or exponentiated coefficients, and which column they are in (often both are included in research papers, but for this purpose I am fine with you only including change in logged odds and then using exponentiated odds in the text).

You should explain what the independent variables' estimated associations with the dependent variable mean. If you are using a large number of additional variables that are not necessarily of substantive interest, but are things you feel like the model should 'control' for (e.g. potential confounders), you do not have to report these.

You should use odds ratios (exponentiated odds) in your interpretation and explanation. For example, "Women were at 1.75 times greater likelihood of being diagnosed with reduced mobility than men (B = 0.56)".

You may wish to include predicted probability plots (e.g. using the `ggeffects` package) for important independent variables to better describe the likelihood of the outcomes based on the independent variables, however, this is not an expectation (it is increasingly a requirement in academic journals though, so can be worth learning). It can be helpful for describing the strength of a continuous predictor (e.g. the linear effect of age on developing a health problem or not), or for ensuring that findings do not sound too "overblown" (e.g. a 8 times increase in the likelihood of something that only happens on average 0.02% of the time means it's still a very unlikely event).

Cluster Analysis

The primary means of reporting your cluster analysis should be through either a cluster biplot/cluster plot (for k-means) or through a dendrogram (for HCA). You should justify why you chose a particular cluster solution (e.g. the number of groups for k-means, or where to cut the tree in a HCA) using an appropriate method (e.g. elbow plot, silhouette plot, or distance for a HCA).

You should then try to 'make sense' of and describe the cluster solution you ended up with based on bivariate descriptive statistics about the cluster membership, the variables used to create the clusters, and other variables of interest (if relevant). Provide the percentage of cases found within each cluster. These descriptive statistics broken down by group should be presented in the form of a table.

Describe the characteristics of each of the clusters and what defines them. For example, "The states in cluster 1 were defined by relatively high vehicle theft and burglary rates, with relatively low murder and drug misuse rates, whereas the states in cluster 2 were defined by high murder and drug misuse rates and low vehicle theft and burglary rates. The states in cluster 3 were defined by high rates of all crimes". You may then want to come up with names or labels for the clusters you found to better lead into your discussion, e.g. "US States could be classified within four general groups according to their crime levels: overall low

rates of all crimes (10% of states); overall high rates of all crimes (20% of all states); high rates of property crime only (60% of states); and high rates of interpersonal crime only (10% of all states)".

Spatial Analysis

The primary means of reporting your spatial analysis, if you choose to use this method, should be through the use of choropleth maps and the Moran's I statistic. These should generally be reported first, and then described for a general audience. You may find it useful to highlight particular sites of interest, especially with the needs of visually impaired audiences in mind.

If you are comparing multiple areas your choropleth maps should be generated separately (unless they happen to be geographically adjacent), and your Moran's I statistics should be collated together in a single table showing the area's name and the associated Moran's I statistic for the variable(s) of interest (however, they should be estimated separately). When explaining your findings, you should interpret what the Moran's I statistic suggests and then describe how this appears to be represented in the choropleth visualisation.

Depending on your research question you may wish to include some additional bivariate statistics (for example, to show the correlation between your variable of interest and other factors) or Moran's I measures for several variables that are related. Additionally, you might also wish to include a multiple linear regression model that explores some of the predictors of the variable of interest. However, this is not an expectation.

Discussion

The discussion is the section of your paper where you are able to explain how your research findings correspond with, add to, or diverge from the existing literature. They are also commonly used to acknowledge the limitations in the research design (or strengths), as well as the reporting of any violated assumptions in the model (while plots that show these violated assumptions are often placed in an appendix, you should include them inline here).

Synthesising your findings with the existing literature

Objective: Demonstrate that you are able to evaluate your research findings against existing literature and context.

- How do your findings corroborate, contradict, or contribute to existing research?
- What might the policy or social consequences of this research be, if relevant? Could these findings be used to identify ways of reducing inequalities, or reducing incidences of a social problem?
- What further research could/should be done? What different data could be gathered to take this research futher?

Acknowledging the limitations of the study design

Objective: Critically reflect on any limitations in study design.

- What were you wanting to do but couldn't?
- What variables may have been not included in or unmeasured in the dataset but are potentially important, especially if they are confounders (omitted variable bias)?
- Does the study design that collected the data limit the findings in any way? E.g. are we unable to generalise to a wider population? Unable to make claims about causality?

Reporting any violation of assumptions

Objective: Demonstrate that you can conduct any relevant checks for violation of assumptions and explain what they mean and how future research might address them.

- Were any of the assumptions made by the method you used violated by the data?

- If so, which assumptions were these and how might them being violated have affected the results you found?
- Is there anything you could recommend future studies do to correct for this violation of assumptions?

Conclusions

Objective: Demonstrate that you are able to summarise the key research contributions from this research paper for a general audience.

Finally, a research paper should return to the research question and rationale that guided it. Consider responses to the the following questions which can form parts of your discussion and conclusions:

- Did the research provide any answers to your research question? If so, summarise what answers it provides. If not, try and explain what might have prevented this.
- What is the one "take-away" message you want a reader to have from this paper?
