SMI606 Introduction to Quantitative Research: Assessment 1

Dr Calum Webb

2023-08-25

Assessment outline

For this assessment you must submit a short report on an original piece of quantitative research conducted in `R` on a topic of your choice using secondary data.

In this report you must: describe the research question being explored and provide a rationale for why you are exploring this research question; describe the data being used; describe (using statistical methods, data visualisation, and in writing) the variables of interest and their association with one another; correctly run, summarise, and interpret the output of a bivariate linear regression model; report on any relevant checks for violation of assumptions; and, in conjunction with a hypothesis test, explain any inferences that can or cannot be made about the general population and about causality in your discussion. You should also briefly conclude with a discussion about how your findings relate back to the research question.

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

Word Limit

1,000 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 (if necessary) are not included in word counts, but in-text citations are.

Assessment Value

30% 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 reports

Below is further guidance for what is meant by each of the aspects expected to be included in the research report. Remember, this report is only 1,000 words in total — which excludes any code, tables, references or data visualisations — so content related to each of the points below should be relatively short. In other words, don't end up writing 500 words just describing the data or you will end up not being able to adequately demonstrate you've met the other learning outcomes. If you gave all of the below sections equal weighting they would only be just over 100 words each. Be brief and concise!

Report Structure: You do not have to structure your research report in exactly the way key learning outcomes are broken up below, nor do you need to use the same chronology. You may choose to combine some parts or further break up others and this may improve the readability of the report.

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.

Research question and rationale

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 report. 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. You do not have to cite literature in this assessment, but it can be helpful to include one or two citations. 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.

Example of a good research question and rationale: *"There has been an increasing emphasis on the relationship between income inequality, poverty, deprivation, and child welfare interventions — rates of children being taken into care or placed on child protection plans (Bywaters, et al. 2019; Keddell, 2020). However, despite being an important part of the lived experience of poverty and deprivation (Lister, 2004; ATD Fourth World, 2019), there has been little focus specifically on the role of housing quality on child protection concerns. This research report will explore whether housing disrepair is associated with increased incidence of child protection plans in neighbourhoods in Northern Ireland."* (Words = ~95)

Note that the research question does not necessarily need to be phrased as a question (with a question mark). The research question here is the final sentence of the example, the preceding sentences are the rationale.

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 of complete population; how the sample was derived (whether it is random or not); and contextual information about the population it is derived from (e.g. is it all from one country? Is it multiple countries? Is it routinely collected as part of a survey? Is it administrative or official statistics?).

Example of a good description of data: *"The Family Resources Survey is a repeated cross-sectional survey of households in the UK published by the Department for Work and Pensions. It includes a stratified, clustered, random representative sample of households and collects data on income, housing, caring, disability, employment, and other factors (DWP, 2021). The 2019-2020 dataset used in this research report included responses from between 19,000 and 20,000 households."* Word count: ~64

Note that this example is clear, brief, and gives all of the necessary details that someone would need to find the same dataset you are using themselves. It also sets the groundwork for what can be said about causality and inference by describing that the sample is random and that the data is from a survey (not a controlled experiment) and is cross-sectional (done at one point in time, not studying the same people over time).

Univariate and bivariate visualisation and descriptive statistics

Objective: Describe (using statistical methods, data visualisation, and in writing) the variables of interest and their association with one another.

You should always describe the variables you are using in answering your research question using both univariate and bivariate statistics and data visualisation (in practice — outside of university assessments — sometimes these details are relegated to the appendices. Regardless, any competent social researcher should produce and inspect them). The visualisations and statistics you choose should be appropriate for the types of variables you are using (e.g. depending on whether they are continuous, categorical, or ordinal).

As well as providing descriptive statistics, **make sure you also explain plainly what is being measured** with some precision, if possible. For example, if you refer to 'child poverty' throughout but do not explain to the reader that child poverty is measured here by the Income Deprivation Affecting Children Index variable, which refers to the proportion of children living in households with income less than 60% of the median, they would not be able to meaningfully interpret some of your results.

Descriptive statistics should be presented in numbered tables (see the following book chapter for how to format tables in `Rmarkdown`, though I don't mind if they are presented as raw `R` output as long as important information is not cut off: https://bookdown.org/yihui/rmarkdown-cookbook/tables.html). Tables should be referred to in text. For example, "*Table 1 presents descriptive statistics for the distribution of the variables used in this study, including their mean, median, mode, standard deviation, and inter-quartile range. The mean value for [Variable X] was [whatever] and 95 per cent of values fell between [this] and [that]...*". You can create nicely formatted tables for many R object types easily using the `stargazer` R package: https://cran.r-project.org/web/packages/stargazer/vignettes/stargazer.pdf

Data visualisations should be presented in such a way that they are clearly labelled and titled with a figure number (e.g. using the `xlab()`, `ylab()`, and `ggtitle()` functions respectively). They should also be referenced in the text and explained. For example, "*Figure 2 shows the correlation between variable [X] and variable [Y] in the form of a scatterplot. The distribution of points in this figure suggests that there is some non-linear association and that the relationship between the two variables becomes weaker at higher values of both...* "

Remember that **you must describe** the statistics and visualisations you use **in the text**. Do not just leave them floating. This is not only good academic practice but makes your work more accessible to those who may have barriers to reading tables or seeing data visualisations. You audience when writing up research is not often people who understand much about statistics, so you need to clarify everything and make it clear (while avoiding giving them a crash course on statistics!). For example:

Don't just write... *The correlation was 0.325.*

Do write... *The correlation between [Variable X] and [Variable Y] was 0.325, indicating a weak positive correlation. As rates of [Variable X] increase, rates of [Variable Y] also tend to increase. This suggests that [things] with higher [whatever] also experience higher rates of [something else].*

Bivariate linear regression

Objective: Correctly run, summarise, and interpret the output of a bivariate linear regression model.

You should present at least one bivariate linear regression in this assessment. Keep this in mind as it means that you should be using two continuous variables. In presenting your bivariate linear regression, you can always use the `stargazer` package to format the output so that it looks nice in an Rmarkdown document (https://cran.r-project.org/web/packages/stargazer/vignettes/stargazer.pdf).

I suggest taking a 'funnel' approach to reporting your results, starting from the least-detailed, most general parts of your model and moving to the most specific:

- Report your R-squared statistic and what this means — what proportion of the variance in your dependent variable is your independent variable able to explain?
- If relevant, report whether the linear relationship between the pair of variables was statistically significant.
- Report whether the relationship between the pair of variables was positive or negative, and interpret what this means.
- Report the standardised coefficient and the strength of this relationship (only if Pearson's R correlation coefficient was not reported earlier).
- Report the exact association between the two variables by means of the intercept and slope values (e.g. for each one-unit increase in X, we expect to see a N-unit increase in Y). Ensure that the unit is defined (what does a one-unit increase mean in context? One standard deviation? Increase of one on a log scale? One percentage point increase?).
- (Optional) Give a contrast of two hypothetical examples with different values of X to illustrate the size of the association, for example:

  A local authority with a [variable X] score of 10 would have a predicted [variable Y] score of 15, whereas a local authority with a [variable X] score of 20 would have a predicted [variable Y] score of 25.

This is more important if you are using variables that have been transformed in some way, for example, logged values or z-scores (changes in standard deviation from the mean). It helps the reader understand the findings in context where they are unlikely to understand, for example, what a z-score increase of 1 means (but this can be converted back to a 'meaningful' amount). For example, if I used z-scores of Income Deprivation Affecting Children (IDACI mean 15, standard deviation 6) in my model, I might compare a local authority with a IDACI 9 to one with IDACI 21. Giving two hypothetical examples is also important for logistic regression models (where you can show predicted probability), so this tip is worth remembering for assessment two.

Testing for violation of assumptions

Objective: Report on any relevant checks for violation of assumptions.

You should report on any tests and visual inspections for violation of linear regression assumptions here, this should include:

- Linearity
- Homoscedasticity/Homogeneity of Variance
- Outliers and leverage points
- Normality

- Independence

For **linearity**, you should include and interpret a plot of residuals versus fitted values and conclude whether there is evidence of any non-linearity. For bivariate linear regression (in this assessment), you may also base this on a scatterplot of the values of the two variables.

For **homoscedasticity**, you should report a spread-location plot and highlight whether the residual variance increases or decreases at different values of X. For bivariate linear regression (in this assessment), you may also base this on a scatterplot of the values of the two variables.

For **outliers and leverage points**, you should include a leverage plot. If your model has any outliers, you should consider comparing the differences in model parameters once these outliers are removed. For bivariate linear regression (in this assessment), you may also base this on a scatterplot of the values of the two variables.

For **normality**, you should report on a Q-Q plot to check that your residuals are normally distributed across values of X. Remember, the assumption is about the normality of the *residuals* of your dependent value for different values of your independent variable, *not about the normal distribution of your dependent variable itself.*

Unless there is an obvious reason why observations may not be independent based on what you know about the dataset (e.g. it's pooled data of the same observations over multiple years), **independence** is not something you would normally need to worry about for this assessment. For multiple linear regression, you would be expected to provide correlation statistics between all predictors and Variance Inflation Factors (VIF).

Remember that violated assumptions are not a death sentence for your research or findings. If you find your research violates and assumption and cannot be corrected (e.g. by transforming a variable to its logged values for non-linearity and non-normal residuals, or by removing an outlier), you should just report what this might mean for any conclusions. The only aspects you are unlikely to be able to "fix" are heteroscedasticity and dependence between predictors.

For models that show heteroscedasticity, this actually tells us something interesting about the data which should be reported. It tells us that our predictions using a linear model are more accurate at some values of X than they are at others. In context, this could actually be an interesting finding (e.g. areas with high poverty don't have worse school results on average, but they do have more variable results).

Remember — you do not need to explain what these terms mean or speak about them at length. Just report on how you checked them, what the results suggested, and make any statements about limitations of or changes to the model if you need to.

Inference and causality

Objective: In conjunction with a hypothesis test, explain any inferences that can or cannot be made about the general population and about causality in your discussion.

As above, and in reference to your description of the data (the kind of sample and data collection method it came from), you should state whether your data can be generalised to a wider population based on a hypothesis test.

In this section you should explicitly show you can interpret a *p-value*, though you do not need to explain what a *p-value* is to a reader. If you *do* feel the need to explain what the *p-value* is, you should avoid the following common misinterpretations (see: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4877414/ for a paper about common misinterpretations):

- "This means that there is less than a 5% chance of finding this result"
- "This means that the results cannot be down to chance"
- "This is the probability that the null hypothesis is true"

- "This means that there is less than 5% chance that the relationship is not real"

A suggested interpretation of a 'statistically significant' result is:

- "An association of this strength is sufficiently unlikely to be observed in a sample of this size if the null hypothesis were the best description of the relationship between the two variables."

A suggested interpretation of a 'non-significant' result is:

- "The relationship we observed was consistent with what we would have expected under the null hypothesis of no association between the two variables at this sample size."

Further, you should make it explicitly clear what can be said about the causality of the relationship you are exploring — including if causality cannot be inferred.

These two points generally come under a 'strengths and limitations' section in a research manuscript; in this, you might also want to write about the limitations in any measurements you might have used or about the risks of ecological fallacy if you are using area-level data.

Discussion & Conclusions

Objective: You should also briefly conclude with a discussion about how your findings relate back to the research question.

Finally, a research report 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.
- Based on the rationale you developed (the argument why your research is of scholarly and/or policy importance), how could this research be used to further our understanding of the topic and/or identify policy implications? How does it corroborate or contradict existing research or theory?
- What were the limitations and what would a good follow-up question for further research be?

Some people think of the conclusions and discussion section as 'tying everything up in a neat bow': making it clear what the link between the start of the report (the context and rationale) and the original research is.
