

SMI606: Assessment 1

Guidance & Walkthrough

Dr. Calum Webb

Sheffield Methods Institute, the University of Sheffield
c.j.webb@sheffield.ac.uk



Assessment Guidance

Assessment guidance is uploaded on Blackboard.

- Pick a dataset and research question
- Write a rationale
- Describe the dataset
- Describe the data using univariate and bivariate statistics
- Estimate a bivariate linear regression and describe the findings
- Check for violations of assumptions
- Make appropriate statements about inference and causality throughout
- Discuss the findings and conclude the report

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 assignment. 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)



Pick and dataset and research question

What is your research about?

- Look in the description of datasets file found on the assessments page *or* find some data you are interested in.
- Try and pick something you are interested in (the assessment will be less painful that way)
- **Remember** that in assessment 1 you are only looking at **two** variables, so your research question will need to be quite precise



Pick and dataset and research question

What is your research about?

- Look in the description of datasets file found on the assessments page *or* find some data you are interested in.
- Try and pick something you are interested in (the assessment will be less painful that way)
- **Remember** that in assessment 1 you are only looking at **two** variables, so your research question will need to be quite precise

Bad:

⊗ "How does poverty and local political party control cause increases in housing insecurity?"

What's wrong:

- **More than one independent variable**, not appropriate for assessment 1.
- **Implies causality**, is it possible to demonstrate with the data and method?
- **Imprecise outcome/variables**, how is 'housing insecurity' measured in the data? Often it is better to use the specific variable's definition and then widen the concept in the discussion.
- No clear **unit of analysis**, individuals, local authorities, countries?



Pick and dataset and research question

What is your research about?

- Look in the description of datasets file found on the assessments page *or* find some data you are interested in.
- Try and pick something you are interested in (the assessment will be less painful that way)
- **Remember** that in assessment 1 you are only looking at **two** variables, so your research question will need to be quite precise

Better:

☑ "Are higher rates of poverty in local authorities associated with higher rates of homelessness applications?"

What's better:

- **Only one independent and one dependent variable**
- **Restricted to association**, not setting such a high bar to demonstrate causality
- **More precise description of variables**
- **Defines the unit of analysis (local authorities)**



Write a rationale

Why should I **care** about your research?

- Now you have come up with a question, explain to me in a sentence or two **why I should care about it.**
- A good rationale has an explicit call to policy discussions or academic debate.
- The rationale allows you to explain how your more precise research question fits within the wider scholarship.



Write a rationale

Why should I **care** about your research?

- Now you have come up with a question, explain to me in a sentence or two **why I should care about it.**
- A good rationale has an explicit call to policy discussions or academic debate.
- The rationale allows you to explain how your more precise research question fits within the wider scholarship.

Bad:

⊗ "Not much is known about how strongly related poverty and housing insecurity is. However, it is important to research this so that we can identify policies to prevent homelessness. In addition, both variables are continuous and suited to bivariate linear regression."

What's wrong:

- **Vague, unevidenced claim about lack of research** - is that really true?
- Vague claim about it being important to research - **why** is it important to research??
- Okay for an undergraduate assignment, but at postgraduate we want you to think more deeply about this.



Write a rationale

Why should I **care** about your research?

- Now you have come up with a question, explain to me in a sentence or two **why I should care about it**.
- A good rationale has an explicit call to policy discussions or academic debate.
- The rationale allows you to explain how your more precise research question fits within the wider scholarship.

Better:

✓ "The end of the COVID-19 pandemic has seen a surge in housing demand across England and Wales, especially in historically deprived cities as internal migration out of London has increased. Costs associated with housing have been one of the greatest contributors to a cost of living crisis; as this crisis continues and a recession looms, understanding the association between homelessness and poverty rates may help policymakers direct resources towards people areas of the country at the highest risk of homelessness."

What is better:

- **Direct** link to contemporary policy concerns.
- Demonstrates the ability to think about how the research findings might result in some impact.



Describe the dataset

Where is your data from? How was it collected, and what can that tell us about the wider population or causality?

- If you are using one of the provided datasets: **look in the description of datasets folder on Blackboard for clues on how it was collected.**
- Who collected the data?
- What is the unit of analysis (country? local authority? neighbourhood? person?)
- Is the data an entire population of interest? A random sample? A stratified random sample? Or an opportunity sample?
 - Based on this, what will we be able to say about inference (Check [Week 4 Slides](#))?
- Does it come from a causal design? Keep this in mind when talking about causality.



Describe the dataset

Where is your data from? How was it collected, and what can that tell us about the wider population or causality?

- If you are using one of the provided datasets: **look in the description of datasets folder on Blackboard for clues on how it was collected.**
- Who collected the data?
- What is the unit of analysis (country? local authority? neighbourhood? person?)
- Is the data an entire population of interest? A random sample? A stratified random sample? Or an opportunity sample?
 - Based on this, what will we be able to say about inference (Check [Week 4 Slides](#))?
- Does it come from a causal design? Keep this in mind when talking about causality.

Bad:

⊗ "The data is the la-dat.csv. It includes 6 continuous variables and 2 categorical variables."

What's wrong:

- **No details that tell us anything about the source**
- **No details on the sample**
- **No details on the unit of analysis**
- **Variable types not important for the reader**



Describe the dataset

Where is your data from? How was it collected, and what can that tell us about the wider population or causality?

- If you are using one of the provided datasets: **look in the description of datasets folder on Blackboard for clues on how it was collected.**
- Who collected the data?
- What is the unit of analysis (country? local authority? neighbourhood? person?)
- Is the data an entire population of interest? A random sample? A stratified random sample? Or an opportunity sample?
 - Based on this, what will we be able to say about inference (Check [Week 4 Slides](#))?
- Does it come from a causal design? Keep this in mind when talking about causality.

Better:

✓ "The data used includes variables about homelessness applications per 10,000 households in 2021, sourced from the Ministry of Housing and Local Government, and about the proportion of children living in low income families in 2021, sourced from the Department for Work and Pensions. This data is at the local authority level and includes all 151 local authorities. These administrative data are collected routinely."

What's improved:

- **Tell us about the source** for the data.
- **Tells us about the unit of analysis**
- **Tells us this is a population, not a sample of a population.**
- **Tells us this did not come from an experimental design.**



Describe the data using univariate and bivariate statistics

Demonstrate that you can use descriptive statistics and data visualisation to describe the two variables you are interested in independently, and their association with one another.

- **Use the tables in the Week 2 slides to identify the appropriate univariate descriptive statistics and data visualisations and their interpretation for your variables, depending on if they are continuous, categorical, or ordinal:**

[Link to slides](#)

- **Use the tables in the Week 3 slides to identify the appropriate bivariate descriptive statistics and data visualisation and their interpretation for your two chosen variables:** [Link to slides](#)



Estimate a bivariate linear regression and describe the findings

Demonstrate you can express the relationship between the two variables in the form of a bivariate linear regression (Week 6)





This part of the assignment refers to Week 6: Bivariate Linear regression. [Link to slides.](#)

In your assignment, you should report:

- The estimated change in the dependent variable for a 1 unit increase in the independent variable (Estimate - for the independent variable)
- The R-squared statistic.
- The p-value of the estimate (if appropriate)

See the specific assessment guidance for a general workflow.

Most common mistakes:

- Dependent variable is on the wrong side of the regression symbol:
 -  Incorrect:
`poverty_rate ~ homeless_rate`
 -  Correct:
`homeless_rate ~ poverty_rate`
- Interpretation uses the wrong scale, e.g. if poverty rate is measured per 100 children and homelessness rate per 10,000 households:
 -  Incorrect:
A 1 percentage point increase in poverty was associated with a 0.1 percentage point increase in homelessness.
 -  Correct:
A 1 percentage point increase in poverty was associated with a 0.1 household per 10,000 increase in homelessness.



Check for violations of assumptions

Critically assess what violations of assumptions there may be and the consequences this might have for your model.

Important: use the slides and cheat sheet for these from Week 6:

- Linearity
- Homoscedasticity
- Outliers and Leverage Points
- Normality of residuals

If you find that any of these assumptions are violated, you should use the cheat sheet to assess the consequences this might have and what you could potentially do to resolve them: [Consequences cheat sheet](#)



Check for violations of assumptions

Critically assess what violations of assumptions there may be and the consequences this might have for your model.

Important: use the slides and cheat sheet for these from Week 6:

- [Linearity](#)
- [Homoscedasticity](#)
- [Outliers and Leverage Points](#)
- [Normality of residuals](#)

If you find that any of these assumptions are violated, you should use the cheat sheet to assess the consequences this might have and what you could potentially do to resolve them: [Consequences cheat sheet](#)

An assumption being violated is not **bad**; it doesn't mean the findings are worthless. Knowing that an assumption has been violated means that:

- 1) We can communicate to a reader which statistical tools we should be more cautious about interpreting (e.g. standard error) and what we might do in future research to improve.
- 2) It often tells us something **interesting** about our data - e.g. if the strength of an association between two things gets weaker at higher values of one, heteroscedasticity.



Make appropriate statements about inference and causality throughout

- It's okay to interpret your p-values **as long as** you make it clear whether they are meaningfully generalisable to a larger population or not.
- Avoid talking about causality when you are interpreting your analysis specifically (unless you feel you can demonstrate causality), but it is okay to speculate about reasonable causal pathways in the discussion/introduction.



Discuss the findings and conclude the report

What is your answer to the research question? How does it contribute to the rationale?

The discussion and conclusion is your opportunity to bring together your original research with the rationale you established at the start. The stronger your rationale was, the easier you will find writing the discussion and conclusion.

- Be concise
- Don't go beyond what your research can tell you
- Explore any limitations or directions for future research
- It's okay to speculate about causality here as long as you are cautious ("may")

Bad:

⊗ "In conclusion, the research found that a 1 percentage point increase in poverty was associated with a 0.1 per 10,000 increase in homelessness rates. This contributes to the literature by showing that there is a link between poverty rates and homelessness."

What's wrong:

- **Repeats** what is already in the findings section too precisely
- **Vague** statement about a contribution to knowledge
- **No discussion** of future research or policy directions.



Discuss the findings and conclude the report

What is your answer to the research question? How does it contribute to the rationale?

The discussion and conclusion is your opportunity to bring together your original research with the rationale you established at the start. The stronger your rationale was, the easier you will find writing the discussion and conclusion.

- Be concise
- Don't go beyond what your research can tell you
- Explore any limitations or directions for future research
- It's okay to speculate about causality here as long as you are cautious ("may")

Better:

✓ "This research project found that there was a strong association between poverty rates and homelessness rates in local authorities in England; while this is not surprising on its own, it reinforces the concern that already disadvantaged communities will be more likely to experience housing repercussions as a consequence of the cost of living crisis. Policies that distribute funding towards areas with existing high levels of poverty may be effective in preventing homelessness. Further research that takes account of the availability of affordable social housing is needed to better understand the association between poverty and homelessness."



Most common mistakes:

- Too worried about getting the technical bits correct that you ignore the quality of your academic writing.



Most common mistakes:

- Too worried about getting the technical bits correct that you ignore the quality of your academic writing.
- Getting the scales wrong when reporting the results (e.g. not everything is on the percentage point scale)



Most common mistakes:

- Too worried about getting the technical bits correct that you ignore the quality of your academic writing.
- Mixing up the dependent and independent variables (in model or interpretation).
- Getting the scales wrong when reporting the results (e.g. not everything is on the percentage point scale)



Most common mistakes:

- Too worried about getting the technical bits correct that you ignore the quality of your academic writing.
- Mixing up the dependent and independent variables (in model or interpretation).
- Getting the scales wrong when reporting the results (e.g. not everything is on the percentage point scale)
- Interpreting the p-value when it's not appropriate, or getting the interpretation wrong ($p < 0.05$ = significant)



Most common mistakes:

- Too worried about getting the technical bits correct that you ignore the quality of your academic writing.
- Getting the scales wrong when reporting the results (e.g. not everything is on the percentage point scale)
- Mixing up the dependent and independent variables (in model or interpretation).
- Interpreting the p-value when it's not appropriate, or getting the interpretation wrong ($p < 0.05$ = significant)
- **Trying to explain how the statistical methods work; what a p-value is, etc.** You are not writing a methods textbook: you are writing a research report as an applied user of statistics.
 - It's better to only write: 'There was a statistically significant association between X and Y ($p < 0.05$)' ✓
 - Than to get it wrong by writing:
 - 'There was no statistically significant association between X and Y ($p = 0.5$), this means the null hypothesis is true' ✗
 - 'There was a statistically significant association between X and Y ($p < 0.05$), which means the results are very unlikely to be due to chance' ✗
 - 'There was a statistically significant association between X and Y ($p < 0.05$), which means there is a more than 95% probability of the association being true' ✗



Most common mistakes:

- Too worried about getting the technical bits correct that you ignore the quality of your academic writing.
- Getting the scales wrong when reporting the results (e.g. not everything is on the percentage point scale)
- Mixing up the dependent and independent variables (in model or interpretation).
- Interpreting the p-value when it's not appropriate, or getting the interpretation wrong ($p < 0.05$ = significant)
- **Trying to explain how the statistical methods work; what a p-value is, etc.** You are not writing a methods textbook: you are writing a research report as an applied user of statistics.
 - It's better to only write: 'There was a statistically significant association between X and Y ($p < 0.05$)' ✓
 - Than to get it wrong by writing:
 - 'There was no statistically significant association between X and Y ($p = 0.5$), this means the null hypothesis is true' ✗
 - 'There was a statistically significant association between X and Y ($p < 0.05$), which means the results are very unlikely to be due to chance' ✗
 - 'There was a statistically significant association between X and Y ($p < 0.05$), which means there is a more than 95% probability of the association being true' ✗

If you think any of the three above is accurate it's not so much that you don't understand the p-value but that it has no shorthand, intuitive definition: that is why we use the shared language of "significance".



Live demo: how I would do Assignment 1

