ISSU0053: Coursework Instructions

Session Two, 2024

1 Introduction

Please read and understand these instructions before you begin the assessment.

The coursework assessment will begin with the release of these instructions on the ISSU0053 course Moodle page within the "Assessment 1 – Coursework – Session Two" section during the session taught on Wednesday 24th July.

The intention of the assessment is for you to apply the techniques you have learned during the first two weeks of the course to a real-world dataset made up of a number of variables measured for subregions of London.

A copy of the data to be analysed is available as a CSV file on the course Moodle page within the "Assessment 1 – Coursework – Session Two" section.

The coursework assessment makes up 50% of your mark for ISSU0053.

2 Data

The data are real measurements recorded for the division of the entirety of Greater London into 983 subregions, known as Middle Layer Super Output Areas.

The majority of the data comes from the census completed in 2011, with other non-census data taken as close to that time as possible.

While it is possible to identify each subregion from the data, you should not bring in data outside of that provided to you (that is, you should not bring in entire new variables; however, you can use research to compare values/results for this data to other data, e.g. comparing summaries of this data to the same summaries for the UK as a whole, if you feel that it strengthens your submission).

Variable Name	Description
MSOA	"Middle Layer Super Output Area" code
Median_HP	Median house price ("house" here and throughout is referring to a dwelling, and so the data includes flats, etc.)
Borough	Borough within which the subregion is located
Inner	Whether the subregion is located within an Inner London or Outer London borough
Area	Area of London (Central/East/West/North/South) within which the borough containing the subregion is location
Political	The Political party in charge of the borough within which the subregion is located (Conservative, Labour, Other – including boroughs where no political party has a majority and boroughs where a third party has a majority)
Pop_Density	Population Density (People/Square km)
Aged_65Plus	Percentage of the population who are aged 65 and over
OnePerson_HH	Percentage of households made up of a single person
Born_UK	Percentage of the population who were born in the UK
Owned	Percentage of houses which are owned (potentially with a mortgage)
Unemployed	Percentage of individuals who are unemployed

Obesity	Percentage of adults who are obese (BMI of at least 30)
Female_LE	Life expectancy at birth for females

3 Submission structure

You should structure your analysis and subsequent write-up according to the below headings.

3.1 Exploratory data analysis

The first step in any data analysis is to explore the data, to get a sense of what the variables represent and the potential for relationships between them.

Your submission should include three separate, distinct exploratory analyses, each of which contains:

- The results of a single numerical calculation (e.g. a summary statistic or the results of a hypothesis test).
- A single plot.
- A discussion (300 words or fewer) of what your numerical result and plot tell us about London.

Note that:

1. Each of your analyses will be marked out of 8 marks (for a total of $8 \times 3 = 24$ marks overall).
2. Marks will be awarded for the degree of insight shown in each part of the analysis. A numerical result and/or plot which is not discussed will receive a poor mark – a large proportion of the marks will be awarded based upon the degree to which your discussion correctly interprets your results and illustrates why they are insightful.
3. Variety across the three analyses will be rewarded. For example, submissions which repeat the same analysis and discussion for three different sets of variables fail to show a breadth of understanding and will receive a poor mark.
4. You are very welcome to initially perform multiple calculations and plots before you arrive at your most insightful pairings. Only these final pairings should then be included in your submission. Your discussion should focus on your final pairings, but can briefly mention the other numerical results and plots you considered in the process of arriving at your final pairings.
5. The R code to produce your numerical results and plots should be included within your submission (first numerical result and plot, followed by associated discussion, followed by second numerical result and plot, and so on).
6. Neither of your analyses should include `Median_HP`, as this is the focus of the next part of the assessment.
7. You are free to transform, combine, and identify and potentially remove any outliers from the data. Any such decisions should be justified in your discussion.

3.2 Regression

`Median_HP` is the focus of this part of the task. How can the other variables be used to understand the variability in and predict the value of `Median_HP` via a regression model?

Your submission should include:

- A principled, methodological approach to obtaining what you believe to be the best regression model for understanding the variability in and making predictions of `Median_HP`.
- A discussion (800 words or fewer), supported by appropriately chosen plots and numerical results, which explains at each substantive step leading up to the identification of your best model what you are doing, why you are doing it, and what the results tell you.
- An interpretation (200 words or fewer) of what your best models tells us about the relationship between `Median_HP` and your chosen covariates, and how accurate we can be in those conclusions.

Note that:

1. This component of the assessment will be marked out of 24.
2. Marks will be awarded for the breadth and accuracy of the understanding and implementation of course material taught up to the end of the second week of the course.
3. Your argument that your final model is the best one is likely to be based both upon quantitative (e.g. prediction accuracy) and qualitative (e.g. the degree to which the assumptions underlying regression are satisfied) metrics. It is completely possible for different students to acceptably put forward different choices for the best model if they are each convincingly argued for. On the other hand, it is possible for two students to put forward the same best model but for one to receive dramatically fewer marks than the other – the focus is more on the process of how you arrive at the best model and the justification as to why you consider that particular model to be the best, explained through your discussion, than it is the exact combination of covariates in your best model.
4. You are very welcome to investigate multiple approaches to identifying your best model. The first part of your discussion can touch upon each of these approaches, with additional focus applied to the approach leading to your best model. The second part of your discussion should focus on interpretation of only your single best model.
5. Your discussions should be supported by reasonable quantities of appropriately chosen numerical results and plots, produced using R code contained in the submission. Excessive quantities of unnecessary plots and/or calculations will be penalised.
6. You are free to transform, combine, and identify and potentially remove any outliers from the data. Any such decisions should be justified in your discussion.
7. The target audience of your discussion should be something like a student at the lower end of the class – one who can recognise the names of methodologies, but would need some direction as to which elements of R output are important, and clear interpretation of those results in the context of the data.

3.3 General

Up to two marks will be awarded for submissions which:

- Are easy to read, with a clear structure and correct use of spelling, punctuation and grammar.
- Include code which is relatively clearly interpretable.

Note that:

1. You are not expected to write anything in your submission under this heading.
2. Submissions which include unintelligible written elements may well both fail to receive these additional marks and also lose marks in the earlier components. Your submission can only be

marked based upon how you express yourself in that one document, so it is important that you take some time to put it together with care.

3. Submissions whose code fails to run on a neutral machine (e.g., that of the marker, or one which could be found in the computer rooms the lecture is taking place in) may well both fail to receive these additional marks and also lose marks in the earlier components. You are encouraged to test run your code on a fresh R workspace on one of the lecture room computers to ensure that it runs before you submit it. If your code does fail to run on a neutral machine, there is a greater likelihood that it can be repaired by the marker and be awarded at least partial marks if it written in a way to make it interpretable to someone other than yourself.

4 Submission

4.1 Submission format

You should submit at least two files. The first should be a completed copy of the Coursework Coversheet, the template for which is available to you on Moodle. The second should be a single .Rmd (R Notebook) file, named as "[your candidate code] Coursework". For example, if your candidate code is ABC123 then your submission should be a single R Notebook file named "ABC123 Coursework". (If you include a bibliography within your submission using a separate .bib file then you should upload that as an additional third file. The bibliography file can be named however you choose.)

Your submission should contain within it your candidate code, but should not contain your name.

4.2 Submission length

Word counts are provided in the sections above outlining what is required for each part of the assessment. These are purely wordcounts for the discussions, after excluding R code chunks and ignoring any words included purely as part of the bibliography.

The word counts are an upper limit, not a guide for how much you are required to submit. If you can clearly explain the breadth and depth of your understanding more concisely then your submission will not automatically be marked lower.

Any submission which is over the permitted length will suffer a penalty of 10 percentage points, although any such penalty will not reduce a mark below the pass mark of 40%.

4.3 Submission deadline

You must complete your submission via the link in the "Assessment 1 – Coursework – Session Two" section of the course Moodle page before the deadline of 9.45am on Tuesday 30th July.

There are standard non-negotiable penalties for late submissions which you can read about in the UCL Academic Manual. Any extension to the deadline can only be granted where a student has a Summary of Reasonable Adjustments (SoRA) or has successfully claimed for Extenuating Circumstances. If you have a SoRA and wish to activate the coursework extension, please contact t.honnor@ucl.ac.uk in good time to arrange this. If you experience Extenuating Circumstances which impact your attempt at this assessment, you are encouraged to contact the central UCL Summer School team as soon as possible to explore options for mitigation.

4.4 Plagiarism, collusion and referencing

By completing your submission, you are agreeing to have read and understood the "Academic integrity, plagiarism and collusion" document within the "Assessment 1 – Coursework – Session Two" section of the ISSU0053 course Moodle page.

References to any sources (including AI tools) should be included using your choice of a standard referencing system.

Submissions will be run through the Turnitin system.

4.5 Use of AI tools

UCL assigns assessments to one of three tiers, depending upon to which AI tools can be used on the assessment. This assessment falls under the second tier: AI tools can be used in an assistive role.

You may use AI tools to support you in completing this assessment. AI tools must not be used to write the assessment for you, any text appearing within your discussion must be written in your own words and not simply copied and pasted from the output of an AI tool.

UCL's guidance on AI tools notes that:

- Before using generative AI, you should ensure that:
    - You understand the limitations and risks of using generative AI.
    - Your assignment/research remains your own work.
- Generative AI can be a useful starting point to gather background information on a topic, but be aware that:
    - Generative AI produces information that may be inaccurate, biased, or outdated.
    - Generative AI is not an original source of information: it reproduces information from unidentified sources.
    - Generative AI may fabricate quotations and citations.
    - It is always best to refer to original and credible sources of information.
- If you do choose to use generative AI tools, you must always:
    - Critically evaluate any output it produces.
    - Carefully check any quotations or citations it creates.
    - Correctly document your use of the tools so that it can be appropriately acknowledged.

5 Queries

Any queries about the arrangements for this assessment should be emailed to t.honnor@ucl.ac.uk. Any queries which require and receive an informative response will be shared with the whole class via the course Moodle page and/or in one of the teaching sessions.