Ast2

Due March 13 at 11:59pm on moodle. No email submissions.

The assignment can easily be done individually but pairs/groups are *encouraged*. Only one group member should submit the assignment. Include all group member names. Other group members will have a moodle warning they didn't do an assignment; ignore that. I manually enter grades for everyone.

You're welcome (actually, you're encouraged) to use online resources to help you find facts/insights for this assignment. This includes using Generative AI. However, you are ultimately responsible for content, and you must *document* which knowledge/tool you use and where. Any content without explicit citation is assumed to be your own work.

To document your work, I require a properly formatted bibliography/references. In-text citations should be author-year and the bibliography itself may be formatted any compatible way. I strongly recommend LaTeX/quarto for this. (Most of you will need to learn LaTeX for your final Master's thesis/project, anyway!) If you'd like information on setting up a LaTex document, I have this information or you can google.

https://github.com/mateenshaikh/trumscdscproposaltemplate

Assignment objectives

- Read/lookup documentation on data
- Perform simple linear regression analysis and construct appropriate visualizations
- Investigate residuals
- Write a well-written, short report

1 Data

You will be analyzing Anscombe's Quartet. It's a default dataset in R, so you can directly access through the variable `anscombe` without calling any library. You can find it online as well. The entire dataset is shown it below.

```
anscombe
```

```
   x1 x2 x3 x4    y1   y2    y3    y4
1  10 10 10  8  8.04 9.14  7.46  6.58
2   8  8  8  8  6.95 8.14  6.77  5.76
3  13 13 13  8  7.58 8.74 12.74  7.71
4   9  9  9  8  8.81 8.77  7.11  8.84
5  11 11 11  8  8.33 9.26  7.81  8.47
6  14 14 14  8  9.96 8.10  8.84  7.04
7   6  6  6  8  7.24 6.13  6.08  5.25
8   4  4  4 19  4.26 3.10  5.39 12.50
9  12 12 12  8 10.84 9.13  8.15  5.56
10  7  7  7  8  4.82 7.26  6.42  7.91
11  5  5  5  8  5.68 4.74  5.73  6.89
```

The dataset are actually 4 pairs of data, `x1` matches with `y1`, `x2` matches with `y2`, etc..

All of the regression model fits are the same, but the residuals are quite different. The thesis of your report will be that for these data sets, the most meaningful analysis is not what the model explains but analyzing what the model doesn't explain (residuals).

1.1 Report Requirements

Find all the parameters of the model from maximum likelihood (up to 2 decimals) and the coefficient of determination. You will have to discuss these.

You will then analyze the residuals of each data set. You can consider the means, variances, histograms, plots of residuals against fitted values, etc.. It's your choice how you approach this.

You'll comment on what you can glean about model fits based on your exploration of the residuals (there's no single right way). You should *not* find a method/quantity/visualization that solves the issue in modelling, you just need to identify *what* the problem is. (Addressing it is later in the course!)

1.2 Report guidelines

Your report should be a maximum of 800 words (less than 3 pages of solid text). This limit excludes figures, tables, and references. Text should be size 12 of a standard font and double spaced with 1 inch margins. There should be no code nor output in the body of the report. The report can have an appendix with any other content you wish (but by definition, your report should be complete without the appendix content). The report should not include any derivations nor non-trivial calculations.

Importantly, every sentence, every table, and every figure must non-trivially and clearly contribute to your analysis in a meaningful way. Simply adding/padding content, overly redundant information, or if I don't see the value of a sentence/table/figure will be penalized for poor style.

1.3 Grading

You will be graded according to a weighted geometric mean according to the following scheme.

Weight	Criterion
10	Adhering to guidelines
45	Correctness
45	Presentation/style