IE 493: Midterm Project

Summary

For the midterm project, you are to write a research report. The report must be uploaded in a PDF format to Blackboard. The report should be approximately 5 - 10 pages long, including code and figures, but not including cover page and references.

Midterm project components:

Part 1: Research question

You will identify a research question that you would like to answer with linear regression by studying a particular data set. This question should have the following attributes,

- the question is clear: it provides enough specifics that one's audience can easily understand its purpose without needing additional explanation.
- the question is focused: it is narrow enough that it can be answered concretely.
- the question is concise: it is expressed in the fewest possible words.
- the question is complex: it is not answerable with a simple "yes" or "no," but rather requires synthesis and analysis of ideas and sources prior to composition of an answer.
- the question is arguable: its potential answers are open to debate rather than accepted facts.

For example, you may consider the question, "When factoring for behavioral differences and the variability of the health of KSA adults, does smoking cigarettes cause a practically significant increase in the rate of cancer?" The relationship between smoking is now widely accepted, but it was established largely by statistical techniques. You should consider how this question satisfies the above criteria when you formulate your own question.

This question will give you context for the methods studied in the course.

Part 2: Exploratory analysis

- In this project, each student needs to choose a real data set to perform basic exploratory analysis on this data in the R language. Suggested sources of public data:
    - UCI Machine Learning Repository: Data Sets
    - Awesome public data
    - Data dot Gov
- The selected data set must have at least **8** Attributes.
- Your data analysis should reflect your research question and should help you gain intuition on the relationships between variables that may affect your conclusions. You should interrogate the data for patterns, multi-modality, correlation between variables, summary statistics, trends, outliers, nonlinear scaling, and any points of interest.

Part 3: Linear regression

Following your exploratory analysis, you should construct a linear regression model in the R language. This linear regression model should involve more than two predictors. You need to discuss the questions outlined below.

Questions set (I):

- If you plot the response variable versus the predictor variables, does the relationship appear to be a linear one?
- Are the predictors correlated with the response and/or each other with α=5% significance?
- Produce a model summary for the linear regression. Can you identify which parameters are significant in the model with α=5% significance and which are not? Which are they, and are they correlated with the predictor? Are these variables strongly correlated with other predictors?
- If you remove a parameter that lacks significance, what do you notice about the differences in the estimated values of the other parameters?
- If you remove a parameter that lacks significance, what do you notice about the p-values for the estimated values of the other parameters?
- Examine the confidence intervals for parameters. If a 95% confidence interval includes 0, how does this relate to the P-value for the estimated value of the parameter?
- How is R2 interpreted at a high level and what does this suggest in relationship to your linear model?

Questions set (II):

- Try to systematically select a model in which all estimated parameters are significant. How can you interpret the estimated parameters for this model?
- What do the signs of the estimated values tell you about the relationships between the predictors and the response?
- For a unit change in one of the predictors, which of these have the largest influence on the response?
- Can you produce a prediction for a new case? What is the predicted value and confidence interval for a new observation with all predictors equal to the mean of the predictors?
- What is the predicted value and the confidence interval for a new observation with all predictors equal to the 90th percentile? What do you notice about the differences between these two predictions?

Conclusion

You will need to discuss your results and summarize them in a thoughtful but concise way. Draw connections between your research question and the relationships observed in the above regression and the exploratory data analysis.

Code requirements

You need to include R code in the RMarkdown document to demonstrate your work. Keep the code and the relevant analysis within the page limit. This means iterating on the notebook and reducing it down to the most important analysis and details. Figures should be captioned, with clear and easy-to-interpret labels and graphics.

Document structure

The document should be structured as follows:

- Section 1: Introduce your topic, discussing your research question and why you have chosen it. You should make clear why the question is interesting and relevant.
- Section 2: Discuss your exploratory data analysis. You should include plots of relationships between variables, including line plots and correlation plots, as well as univariate summary figures such as histograms.
- Section 3: Apply a linear model in R that models the relationship of interest in your research question. Discuss the questions above. You are required to answer Questions set (I) and Questions set (II).
- Section 4: Discuss your conclusions as described in the component above.
- Section 5: Works cited.

Learning outcomes

Upon completing the report, the student will demonstrate:

- the ability to formulate an actionable research question;
- the ability to independently acquire new skills in the programming language R;
- the ability to produce an exploratory analysis of statistical relationships in a real data set; and
- the ability to draw connections to the research question from the investigation.

Rubric

The rubric below describes the necessary work delivered per category and associated points in this assignment for full credit. Reports that do not address all of these points, or give inadequate attention to these points will receive partial credit. Adequate attention is contextual and subjective, based on the problem itself and the overall work performed in the report. Students are encouraged to discuss their report in a rough draft with the instructor to get feedback on how to better address these points. Additionally, reports that do not follow document outline, do not use clear language, have formatting or writing errors, or unprofessional figures may be penalized for some of the points below.

Category	Expected results	Total points
Research question	The student effectively discusses their research question, demonstrating the attributes described above. The student clearly describes why this question is relevant and interesting.	10 points
Exploratory analysis	The student effectively discusses connections between their research question and the summary statistics and frequency distributions of the data. The student evaluates the data for the presence of outliers, multimodal and / or skewed distributions. The student makes effective use of plots to demonstrate relationships.	10 points
Regression	The student effectively discusses all required questions.	10 points
Conclusion	The student effectively summarizes their work in the project and draws final connections between the statistical relationships observed and the research question posed at the beginning. The student provides a short discussion on the next steps they would pursue, given more time to study the problem.	10 points
Grand total	All of the above	40/40 points

Due date:

The report, R file, and data file must be submitted on Blackboard by 11:59 PM, Monday, July 5th.