

Linear Regression and Modeling - Project Rubric

IMPORTANT: If the analysis is completed using software other than R, or not written up using R Markdown, the project should receive a 0 regardless of its content.

Part 1: Data (3 points)

- 2 pt for correct reasoning for generability – Answer should discuss whether random sampling was used. Learners might discuss any reservations, those should be well justified.
- 1 pt for correct reasoning for causality – Answer should discuss whether random assignment was used.

Part 2: Research questions (3 points)

- Should be phrased in a non-causal way (1 pt)
- Should be well defined / not vague (1 pt)
- Is clear why this is of interest to the author / audience (1 pt)

Part 3: EDA (10 points)

- 3 pts for plots
 - Plots should address the research questions (1 pt)
 - Plots should be constructed correctly (1 pt)
 - Plots should be formatted well – size not too large, not too small, etc. (1 pt)
- 3 pts for summary statistics
 - Summary statistics should address the research questions (1 pt)
 - Summary statistics should be calculated correctly (1 pt)
 - Summary statistics should be formatted well – not taking up pages and pages, etc. (1 pt)
- 4 pts for narrative
 - Each plot and/or R output should be accompanied by a narrative (1 pt)
 - Narrative should interpret the visuals / R output correctly (1 pts)
 - Narrative should address the research question (2 pts)

Modeling (20 points)

Develop a multiple linear regression model to predict a numerical variable in the dataset. The response variable and the explanatory variables can be existing variables in the dataset, or new variables you create based on existing variables.

- Specify which variables to consider for the full model (1 pt)
- Reasoning for excluding certain variables (2 pts)
- Reasoning for choice of model selection method (2 pts)
- Carrying out the model selection correctly (5 pts)
- Model diagnostics (5 pts)

- Interpretation of model coefficients (5 pts)

Prediction (5 points)

Pick a movie from 2016 (a new movie that is not in the sample) and do a prediction for this movie using your the model you developed and the `predict` function in R. Also quantify the uncertainty around this prediction using an appropriate interval.

- Correct prediction (2 pts)
- Correct quantification of uncertainty around this prediction with a prediction interval (1 pts)
- Correct interpretation of prediction interval (1 pt)
- Reference(s) for where the data for this movie come from (1 pt)

Conclusion (3 points)

A brief summary of your findings from the previous sections **without** repeating your statements from earlier as well as a discussion of what you have learned about the data and your research question. You should also discuss any shortcomings of your current study (either due to data collection or methodology) and include ideas for possible future research.

- Conclusion not repetitive of earlier statements (1 pt)
- Cohesive synthesis of findings that appropriate address the research question stated earlier (1 pt)
- Discussion of shortcomings (1 pt)

Overall (6 points)

- Uploaded HTML document resulting from the Rmd template: 1 pt
- Organization: 1 pts
- Readability of the text: 2 pts
- Readability of the code: 2 pts