STAT 331 – Final project
Prof. Samuel Wong – Fall 2020
This version: November 7

1. Overview

- The project is due **Friday, Dec 11 at 11:59pm** and is worth 40% of your final grade.
- The project will be done **individually**.
- Your submission will consist of two parts:
  - A typed report **maximum 8 pages in length** inclusive of figures. The code used to produce your model and analysis must be shown in an Appendix, which does not count towards the page limit.
  - A text file containing your predictions of the response variable for the given prediction challenge dataset.
- **Lateness penalty** is 10% per day.

2. Dataset description and objectives

Ever since computers were invented, scientists have tried to predict the 3-D structures of proteins based on their amino acid sequences coded by the underlying DNA: this is often referred to as the famous "protein folding problem". Proteins carry out vital functions in our bodies, from blood transport (hemoglobin) to digestion (pepsin). On the other hand, proteins also play a key role allowing disease-causing viruses to be infectious. In this project, you will look at a subset of data from a specific spike protein associated with COVID-19, that enables the virus to attack human cells. Much research around the world has been devoted to better understanding this protein in the past 10 months. A goal of this project is to give you an opportunity to apply linear modeling to this sort of real-world prediction problem.

The file `protein-train.csv` contains 1946 samples of computer-generated structures for the COVID-19 spike protein. The response variable of interest is `accuracy`, which is a measure of how close that computer-generated structure is to a known benchmark structure. You are provided with 685 explanatory variables, and your task is to build the best possible model to predict `accuracy` from these variables. A brief summary of the available explanatory variables:

- `angles`: A score based on the configuration of angles in the structure
- The other 684 variables have names in the form `atomtype1_atomtype2_distance`, and are integer counts of various atomic statistics from the structure. To give an example of how to interpret the variable names, `carbonylC_bbCA_medshort` counts how many pairs of `carbonylC` and `bbCA` atom types are in the structure at a medium-short (`medshort`) atomic distance. In general:
  - `atomtype1` and `atomtype2` will each come from the list: *carbonylC, carboxylC, aliph1HC, aliph2HC, aliph3HC, aromaticC, scTrpN, scHisN, scAGN, scLysN, scArgN, bbProN, hydroxylO, carbonylO, carboxylO, sulfur, bbN, bbCA, bbC, bbO.*
  - `distance` will come from this list of distance keywords which range from "very short" to "very long" distances: *vshort, short, medshort, medlong, long, vlong*
  - **Note**: Not every combination of `atomtype1`, `atomtype2` and `distance` are in the list of variables given. (Some combinations are very rare and I pre-filtered these from the dataset to keep the number of variables to a more manageable size.)

The file `protein-test.csv` contains a further 1946 samples of computer-generated structures for the COVID-19 spike protein. The same 685 explanatory variables are in this dataset. However, the response variable (`accuracy`) is not provided in this dataset, and your task is to use your fitted model to predict its values in this dataset as best as possible.

Some further details about the data. To create the `protein-train.csv` and `protein-test.csv` files, I first generated a total of 3892 sample structures using a protein folding algorithm. I then randomly allocated half of the samples to `protein-train.csv` and the other half to `protein-test.csv` (with the response `accuracy` hidden). Therefore, this is a so-called "blinded" prediction challenge: I will assess your predictions by comparing them with the actual values in the original dataset that are unknown to you.

3. Project components

A. Report (75% of project grade)

You will submit the written report via Crowdmark (maximum 8 pages long in 12 point font with standard 1-inch margins and single-spaced), which should include the following sections:

- ***Summary***. This should be one paragraph long. Describe the objective of your analysis, an overview of your methods and models, and summarize your main results.
- ***Exploratory analysis of dataset***. As discussed in class, it is always a good idea to look at your data, for example via scatterplots, histograms, summary statistics, etc. before fitting any models. What did you learn or notice about the dataset from these preliminary analyses? Use this section to present and comment on any findings that you think are important or interesting.
- ***Methods***. This section describes what analyses and techniques you used to obtain your final model. Be sure to address these questions in your Methods section, and provide appropriate reasoning and justification for your choices (since they will necessarily involve your subjective judgment):
  - What form of model did you choose? For our course, your basic model will be multiple linear regression. If you decided to use any transformations or other extensions to the basic MLR model, describe them here.
  - How did you decide whether a given model was good? (e.g., model selection criteria)
  - How did you choose which variables to include in (or exclude from) your model? (e.g., search strategy)
- ***Results and discussion***. This section describes the results you obtained by carrying out your proposed methods. Also, discuss anything you learned about the relationship between protein folding accuracy and the given explanatory variables, in the process of your analysis. Here are some questions you should answer in the Results section:
  - Summarize what you found as you applied your methods in obtaining your final model. Which (and how many) models were tested by your selection procedure? Can you make a table or figure to summarize them? How many variables are in your final model and how are they related to the response? What is their statistical significance and interpretation? You do not have to show/interpret every regression coefficient in your main report, but at least give a couple examples. (If you have many variables in your model, you could provide the full list of variables and their parameter estimates via a supplementary table in the Appendix.)

  - Overall, how well does your final model fit the data? Are the regression assumptions reasonably satisfied by your final model? (e.g., residual analysis). Keep in mind, predictive performance is the most important consideration in this problem, so you will not be penalized for potential assumption violations, though you should comment on them here, if any.
  - What do you expect the mean squared prediction error (MSPE) to be, if you were to apply the fitted model on new data? How can you be confident that your model should generalize well to doing prediction on new data?
  - Any other interesting findings? E.g., would you conclude that certain distances and/or atom types more useful for prediction than others?

- ***Appendix***. Provide all of your code in the Appendix. You may also include any supplementary analyses or figures that support your findings in the main report. The appendix does not count towards the page limit.

The grading of your report will consider the following criteria:

- The report is well-written.
  - All required content is included.
  - Ideas are clearly expressed, and written in complete sentences.
  - The most relevant results and findings are shown and discussed in the report (remember that optionally, any supporting analyses or results you wish to show can be included in the Appendix).
  - Subjective decisions (in particular, your choice of methods) are reasonable and well-justified.
  - Results presented are correctly interpreted and insightful.
- The report is well-presented.
  - Appropriately use section and subsection headings to organize your report.
  - The report is within the page limit.
  - All tables and figures have informative captions and are numbered.
  - Figures have legible and informative labels.
  - All tables and figures are properly referred to in the text of your report.
  - Use an appropriate number of significant digits (usually three to four).

B. Predictions (25% of project grade)

You will submit a text file with your predictions of the response variable (`accuracy`), one value per line, corresponding to the 1946 samples in `protein-test.csv`. Upload location and instructions to be announced later. Using your predictions, I will assess the quality of your model by calculating your root-mean-squared prediction error, i.e.,

$$RMSPE = \sqrt{\frac{1}{1946}\sum_{i=1}^{1946}(\hat{y}_i - y_i)^2}$$

where $\hat{y}_i$ is your predicted value for the i-th sample in `protein-test.csv` and y_i is the actual value. The lowest RMSPE in the class will receive a score of 100% for this component. The tentative formula for your grade on this component is

$$\frac{\text{lowest RMSPE in the class}}{\text{your RMSPE}} \times 100\%$$

4. Questions?

Please post any questions you have concerning the project on Piazza. I will compile the answers into a FAQ document which I will keep updated for the class.