Statistics 149 — Spring 2021 — Course Project

Mark E. Glickman

Key information and milestones:

Decision to work alone or as a team: 10:00pm on Friday, April 2, 2021

Written report: Due 10:00pm on Wednesday, May 5, 2021

General description:

The final project involves the analysis of a data set that you are being provided in which you will build a statistical model and then write a report summarizing your work. The report should be no more than 6 pages of text summarizing how you approached analyzing the data, how you made various modeling choices, and the substantive conclusions of your modeling efforts. Projects are to be carried out either on your own, or in groups of up to four students.

Initial steps:

If you want to carry out the project with other people, you should begin the process of identifying other students with whom you want to work. I would like this process to be completed at latest by Friday, April 2. Once you have settled on the group of students involved in the project, you can form the group by going to the People tab on Canvas, select Project Groups, and then add yourself to an empty group project (or one that already has your teammates). If you are working alone, do not join a Project group.

Data description:

The goal of this project is to use the modeling methods you learned in the course (and possibly an occasional application of related methods) to analyze a data set on engagement with a professional society. Members of the American Statistical Association (ASA) have the option of joining "sections", which are subject-area and/or industry-related subdisciplines of statistics. Example sections include the Section on Statistics in Sports, the Health Policy Statistics Section, and the Section on Risk Analysis. ASA members can join as many sections as they wish, or not join any section. Typically the decision to join a section is made when a member joins or renews their ASA membership. A list of ASA Sections can be found at https://www.amstat.org/asa/membership/Sections-and-Interest-Groups.aspx. Please ignore "interest groups" which can be thought of as provisional sections.

A data set was created to investigate factors predictive of being a member of an ASA section. The data set, called `member-data-2020-stat149.csv` and which resides in the Data Sets folder on

Canvas, can be accessed here. The data set contains information on 17,594 ASA members as of March, 2020. The 37 variables in the data set are as follows.

`AnySection` ("Yes" if member is a member of at least one section in March 2020, "No" otherwise)

`JSMtot` (number of Joint Statistical Meetings, the main annual statistics conference for members, attended between 2015 and 2019)

`USA.CAN` ("Yes" if the member resides in the USA or Canada, "No" otherwise)

`DontPublish` ("Yes" if the member requested not to publish their contact information, "No" otherwise)

`MEMTYPE` (Membership type as a categorical variable - see Table 1 below)

`P.SEC.BE`, ..., `P.SEC.TSHS` ("Yes" if the member belonged to the specified section (27 sections possible) at some time prior to March 2019, "No" otherwise - see Table 2 below)

`Age` (age in years of ASA member in March, 2020)

`AgeJoinedASA` (age in years of ASA member when they joined the ASA)

`Gender` (listed as M or F)

`EmploymentCategory` (one of six different employment categories)

`InChapter` (whether the member was a current "chapter" member – see below)

An ASA member may belong to an ASA "chapter" which is a sub-organization of the ASA defined geographically. For example, there is a Boston Chapter of the ASA.

Table 1 displays the 14 different ASA membership types. As part of your analyses, you may want to consider collapsing categories into meaningful groupings especially when categores are similar in meaning and contain very few members.

The correspondence between the 29 abbreviations in the variable names and the full section names are written out in Table 2. Note that the LIDS (Lifetime Data Science) section was new as of 2019, so that `P.SEC.LIDS` had all "No" entries and was therefore removed from the data set. The `P.SEC.RISK` was omitted from the data set for unknown reasons.

Your goal is to model the probability that an ASA member in March 2020 was a member of at least one section (`AnySection = "Yes"`). Addressing this goal successfully can provide the ASA with insights into the variables that relate to engagement with the ASA, and also suggestions for groups of members to target to increase engagement in sections. Using the provided data set, develop a model for membership to any section from the given predictor information. Make sure you explain how you arrived at your final model, provide appropriate model diagnostics, and provide an inferential summary of important features in your final model.

Abbreviation	Membership Type
I2YC	Community College Educator Membership
ICREP	Corporate Representative
IDEV	Developing Country Resident Membership
IFAM	Family Membership
IFREP	Institutional Representative – Faculty
IK12	K-12 Teacher Membership
ILIFA	Life Membership – Active
ILIFF	Life Membership – Fellow
ILIFR	Life Membership – Retired
IPGRD	Early Career Membership
IREG	Regular Membership
ISEN	Senior Membership
ISREP	Institutional Representative – Student
ISTU	Student Membership

Table 1: Abbreviations of ASA Membership types used in data set.

The modeling task will almost certainly require the following elements:

- Identifying variables that are categorical versus quantitative.
- Exploratory analyses of the data to help inform modeling decisions.
- Addressing the presence of missing data.
- Making reasonable choices for the modeling the response, and for the way in which the predictors are included (e.g., on their original scale or possibly transformed to another scale).
- Iteratively improving your model results, possibly through different modeling decisions (e.g., considering different interactions).
- Making sense out of the relationships you find between the response and the predictors.

Instructions for written report:

The main purpose of the written report is to explain the process of analyzing the data, the logic you followed that led to investigate different modeling choices, and substantive conclusions you learned as a result of the modeling task. The text should be no more than 6 pages of text. The six pages can be single-spaced if you like, but please use a font size no smaller than 11pt. You are encouraged to include graphical and tabular summaries where appropriate (these do not count against the 6 pages of text) which can be included as an appendix. Attaching code is not necessary, but you may find it helpful to insert an occasional code chunk if it helps illustrate particular analyses you performed.

You are free to write the report as you wish, but one way to organize the written summary is in the following manner:

- Have your introduction lay out a description of the problem and its importance, a description of the data you are analyzing, and the goals of the project.
- A follow-up section can summarize the results of exploratory analyses of the data, describing key features of the data that will be relevant for your modeling efforts. Any data pre-processing steps should be discussed here.
- At this point, you should describe the main models that you considered, and your decisions to consider alternative models. This section could finish with an explanation of the model or models that you feel are adequate in relating the response to the predictors.
- Based on the modeling from the previous section, you should describe the substantive conclusions that would be important for a non-statistician to understand. You may want to consider and report any pragmatic advice or conclusions that result from the analysis.
- The report can conclude with not only a brief overview of your work, but also a critical evaluation of your overall approach. What aspects of your modeling attempts did you expect would substantially improve inference but did not achieve the desired outcome? What limitations can you offer about your results, or the process that led to your results?

Project Grade

The project is worth 25% of your final course grade. If working on a team, all team members will receive the same project grade. From a grading perspective, the main criteria for a successful project include

- Evidence that you have learned material taught in the course. While you are encouraged to try occasional statistical methods beyond those taught in the course, you should emphasize your experience using tools, methods, and concepts taught in Stat 149, and incorporate them into your report.
- Evidence that you have put some time and thought into the project. It is important, in particular, to demonstrate that you have reflected on the substantive results of the modeling analyses, not just that you have mechanically applied methods without considering the meaning of the results. Avoid rushing through the project as this will produce a sloppy report.
- Clarity of your written summary and correctness of the content. When writing your summary, you should make sure your explanations are clear, and that you are using correct notation and terminology in describing your modeling and methods used. Your notation and terminology should be consistent with that developed in Stat 149 this semester, not with another course that used different notation.

Abbreviation	Full Section Name
BE	Business and Economics Statistics
BIOM	Biometrics
BIOP	Biopharmaceutical
CNSL	Consulting
COMP	Computing
EDUC	Statistics and Data Science Education
ENVR	Environment
EPI	Epidemiology
GOVT	Government Statistics
GRPH	Statistical Graphics
HPSS	Health Policy Statistics
LIDS	Lifetime Data Science (ignored for this analysis)
MDD	Medical Devices and Diagnostics
MHS	Mental Health Statistics
MKTG	Statistics in Marketing
NPAR	Nonparametric
QP	Quality and Productivity
RISK	Risk Analysis (ignored for this analysis)
SBSS	Bayesian Statistical Science
SDNS	Statistics in Defense and National Security
SGG	Statistics in Genomics and Genetics
SI	Statistics in Imaging
SIS	Statistics in Sports
SLDM	Statistical Learning and Data Science
SOC	Social Statistics
SPES	Physical and Engineering Sciences
SRMS	Survey Research Methods
SSPA	Statistical Programmers and Analytics
TSHS	Teaching of Statistics in the Health Sciences

Table 2: Abbreviations of ASA Sections used in data set variable names.