Modeling ASA Section Membership

Blake Bullwinkel, Connor Capitolo, Teresa Datta, Yingchen Liu

May 5, 2021

Introduction

The American Statistical Association (ASA) is a group of over 19,000 professionals in academia, government, research, and business who "promote the practice and profession of statistics, envisioning a world that relies on data and statistical thinking to drive discovery and inform decisions." With these objectives in mind, it's no surprise that the ASA is interested in expanding both its membership and engagement. One of the key ways in which members engage with the organization is through "sections," which are subject-area and industry-related sub-disciplines of statistics. ASA members have the opportunity to join one or more sections, which encompass wide-ranging areas from Statistics in Sports to Mental Health Statistics. [1]

By building models that predict whether or not an ASA member belongs to at least one section, we are able to better understand which predictors are associated with members belonging to sections or not. These insights might inform how the ASA could boost engagement within specific subgroups that are less likely to be actively engaged in sections, for example through targeted marketing.

1 Data

The ASA collected a data set during March 2020 comprising 37 different variables that provide information on 17,594 members. The three numerical variables are `Age` (age of ASA member in March 2020), `AgeJoinedASA` (age of member when they joined the ASA), and `JSMtot` (number of Joint Statistical Meetings attended between 2015 and 2019). In addition, the data set includes 27 binary variables which indicate whether or not an ASA member belonged to a specific section at some time prior to March 2019 (`P.SEC.BE`, ..., `P.SEC.TSHS`). Finally, the data set has six other categorical variables: `USA.CAN` ("Yes" if the member lives in USA/Canada, "No" otherwise), `DontPublish` ("Yes" if member agreed to publish their contact information, "No" otherwise), `MEMTYPE` (14 categories of membership [2]), `Gender` (male or female), `EmploymentCategory` (one of six different employment categories), and `InChapter` ("Yes" if the individual is a current chapter member, "No" otherwise). Finally, the response variable we will be modeling is `AnySection`, which is a binary variable that indicates "Yes" if the member was part of at least one section in March 2020, and "No" otherwise.

1.1 Exploratory Data Analysis

We first wanted to understand the relationships between the response and the predictors. In Figures C.1 and C.2, we plotted the distributions of numerical and categorical variables, respectively, colored by the response variable `AnySection`. As shown in Figure C.1, we found that `JSMtot` and `Age` had different distributions across the two levels of the binary response, while the distribution of `AgeJoinedASA` did not appear to show a difference. As can be seen in Figure C.2, we found that for `Gender` and `EmploymentCategory`, the missing value ("NA") levels were strongly associated with members not belonging to any section. Perhaps this is

[1] The full list of sections can be found at https://www.amstat.org/asa/membership/Sections-and-Interest-Groups.aspx

[2] Detailed ASA membership options can be found at https://www.amstat.org/ASA/Membership/Membership-Categories-and-Fees.aspx

because members who don't take the time to fill out this information are likely less engaged in the ASA in general.

The reverse phenomenon can be observed for `InChapter`, which suggests that members who belong to a chapter are also more likely to be involved in at least one section. In addition, we see that some categories of `MEMTYPE` contain very few members, which prompted us to think about how to collapse the categories into larger groupings (see section 1.4). Another noteworthy finding stems from the correlation matrix in Figure C.3, which shows several strong positive correlations, including between `InChapter`/`JSMtot` and the response, as well as between `Age` and `AgeJoinedASA`. It is important to note that we used the Pearson correlation coefficient, which is typically used for continuous variables, to calculate the correlations.

To gain a better sense of the 27 binary variables that indicate whether ASA members belonged to a specific section at some time prior to March 2019 (`P.SEC.BE`, ..., `P.SEC.TSHS`), we calculated column-wise and row-wise summations of these features. Figure C.4 (row-wise sum) shows the distribution of the number of sections members belong to on a log scale. We found that members belonged to around 1-2 sections. The column-wise summation of these 27 binary variables, as seen in Figure C.5, shows that `P.SEC.COMP`, `P.SEC.GRPH`, and `P.SEC.BIOM` have many members, perhaps indicating that these are popular research areas. On the other hand, sections such as `P.SEC.SI`, `P.SEC.MDD` have fewer than 50 members.

1.2 Missing Data

A number of predictors contain missing values. First, given that there were only 12 missing values for `USA.CAN` (0.068% of the data set), we simply used mode imputation for this variable. Other features, however, had many more missing values: `Gender` had 2,834 (16.1% of the data set), `EmploymentCategory` had 4,216 (24.0%), `AgeJoinedASA` had 3,400 (19.3%), and `Age` had 3,399 (19.3%). Given that the proportion of missing values for each of these variables is greater than 5%, we cannot simply use mean imputation and instead must assume "ignorable missingness" or a lack of relationship between the propensity for these features to be missing and their potential values.

We also examined whether missingness between different features carried any correlation. In Figure C.6, the numbers on the right side of the matrix indicate the conditions (no variables missing for a given row counted as 0, only one variable missing for a given row counted as 1, and all variables missing for a given row counted as 2; the red color indicates which variable is missing), the numbers on the bottom indicate the total number of rows that are missing for the given feature, and the left hand side of the matrix is the number of missing values that correspond to the conditions on the right. As we can see, the missing values of `Age` and `AgeJoinedASA` are systematic because there is significant overlap in missing values for the two predictors (3,398 of the rows both have missing values).

In light of this, we decided to use the `na.convert.mean()` function, which for numerical variables replaces missing values with the mean and adds a treatment contrast-coded binary variable to indicate whether the value was missing or not (this allows us to understand whether a missing value contributes to the linear predictor beyond the mean of the observed value). For categorical variables, it simply adds an additional "NA" level.

1.3 Outliers in Numeric Data

Several values of the `Age` and `AgeJoinedASA` numeric variables were suspiciously low or high. For example, one member's age when they joined the ASA was listed as 115, while their age was listed as 48 in March 2020. These values are contradictory and suggest that other inaccuracies may exist in the data, perhaps due to user input errors or other administrative issues with the data collection process. Without full knowledge of that process, we decided to simply remove individuals who were at the extremes of the `Age` and `AgeJoinedASA`. We felt it was reasonable to remove individuals below the age of 16 and above the age of 100 from the dataset, as well as individuals who joined the ASA above the age of 90 and below the age of 16. This reduced the dataset size from 17,594 to 17,577 ASA members, which is a decrease of only 17 data points (0.00097% of the data set).

1.4 Membership Type Groupings

Given the small number of members who have certain membership types, we also decided to collapse the `MEMTYPE` predictor into more meaningful groupings that might provide more informative results. As shown in Figure C.2, `MEMTYPE` has 14 levels with highly varying densities, and our exploratory analysis revealed that three membership types occurred fewer than 25 times in the entire data set. Initially we considered grouping these membership types based on intuitive similarities, for example grouping together education or lifetime-related types. However, we ultimately decided to take advantage of potential correlations between membership types and the `AnySection` response by using the `regclass` R package, which automatically suggests level groupings. Figure C.8 shows the segmentation between membership types according to their corresponding `AnySection` values. Specifically, four groupings were created: group A (IFREP, ILIFR, ISREP), B (ICREP, IK12), C (IDEV, ISTU), and D (I2YC, IFAM, ILIFA, ILIFF, IPGRD, IREG, ISEN), which we added to a new categorical variable `MEMTYPE_new`.

2 Modeling

After pre-processing and cleaning the data as described above, we obtained a final data frame that was used for the remainder of the analysis. In order to compare the predictive accuracies of the various models considered in this report, we also split the data into train and test sets using a 75-25 split.

2.1 Binary Response GLM

Given that our goal is to model the probability that an ASA member in March 2020 was a member of at least one section, we started with a standard binary response GLM with a logit link function. Due to the large number of features, we incorporated predictor variables by pre-specifying eight models and performing likelihood ratio tests in order to decide whether a smaller model could be rejected in favor of a model with additional predictors. The deviance values for all models tested are shown in Table 3 (Appendix B).

First, we note that only one of the missing indicators for `Age` and `AgeJoinedASA` should be included in our model, due to the high correlation between these two variables, as elaborated above and shown in Figure C.6. In the modeling that follows, we used only the `Age` missing indicator. Next, we tested for the significance of predictors excluding the section indicator variables (`P.SEC`) using nested combinations of features in Models 1-3 (see Appendix A for full model specifications). Using likelihood ratio tests, Model 1 was rejected in favor of Model 2 with $p = 2.2\text{e-}16$, and Model 2 was also rejected in favor of Model 3 with the same p-value.

Models 4-6 built upon Model 3 by adding different subsets of the section indicator variables, which were determined based on broad categories related to medicine/health, statistical theory/methods, and other (see Appendix A). Model 7 contains all section predictors and was preferred over Model 3 with $p = 5.9\text{e-}06$. Further, Models 4, 5, and 6 were rejected in favor of Model 7 with $p = 0.027$, 9.6e-06, and 1.3e-05, respectively. We therefore prefer the model with the full predictor set, excluding the missing indicator for `AgeJoinedASA`.

Finally, we considered interaction terms. We suspected that the age of ASA members may be related to their employment category and the number of Joint Statistical Meetings attended between 2015 and 2019. Therefore, Model 8 incorporates interactions between `Age` and `EmploymentCategory` as well as `Age` and `JSMtot`, and was preferred over Model 7 with $p = 7.39\text{e-}06$. Based on this analysis, we concluded that the preferred GLM is Model 8. In order to assess our model fit, we ran several standard GLM diagnostics. On the left of Figure 1, we see that the averaged residuals are small and evenly scattered about the zero line. The absence of residuals below zero for averaged fitted probabilities greater than around 0.9, however, might suggest minor non-linearity. The plot on the right of Figure 1 shows that all the Cook's distances are well below 1.0, suggesting that no observations are overly influential.

Despite the lack of issues observed in these diagnostic plots, running a Hosmer-Lemeshow test with 10 groups has an associated p-value of 0.0023, which leads us to reject the null hypothesis at the 0.05 significance level and indicates that our model has a lack of fit. The ratio of residual deviance to residual degrees of freedom is $15003/13129 \approx 1.14$, which is close to 1 and suggests that our model most likely does not suffer from

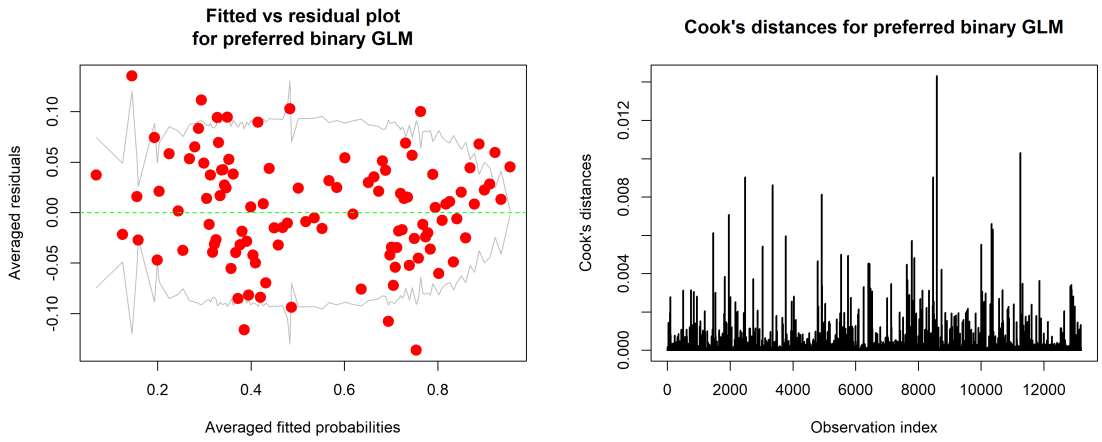

Figure 1: Model diagnostics for preferred binary response GLM

overdispersion. However, there may exist non-linear relationships between the predictors and the response, which motivates us to try a generalizaed additive model (GAM).

2.2 Binary Response GAM

To address the issues with the GLM highlighted in the previous section, we tried fitting a GAM using the same set of predictors as the preferred GLM (Model 8) with smooths for `Age`, `AgeJoinedASA`, and the interactions with `Age`. All other predictors were incorporated linearly. Comparing the GLM and GAM models, we found that the GLM had a residual deviance of 20,073, while the GAM a residual deviance of 19,954. A likelihood ratio test yielded a p-value of 2.2e-16, indicating that the GAM is strongly preferred over the GLM. This also suggests that the contribution of at least one of the predictors is significantly non-linear. Figure 2 plots the smooth of `Age`, which follows a clear parabolic shape and allows us to visualize its non-linear relationship with the response. The smooth of `AgeJoinedASA` also suggests that this predictor has a non-linear relationship. While GLMs enforce a linear combination of features, GAMs allow any smooth function (e.g. quadratic-like as seen in the smooths) to be applied to the predictors.

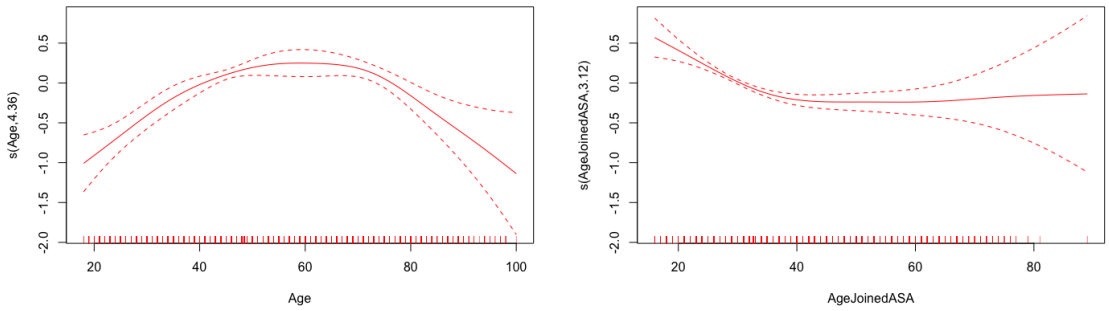

Figure 2: Final GAM smooths

Given that the GAM is strongly preferred over the GLM modeled in Section 2.1, we take the GAM as our final model and summarize the coefficient estimates for the statistically significant linear predictors in Table 1. Note that the entire data set (train and test) was used to generate these estimates, and the 95% confidence intervals shown in the final column were calculated using Wald intervals. While this is a slightly crude approach, calculating confidence intervals via profile likelihoods for GAMs is currently not supported by the `confint` R function. See Section 3.1 for our interpretations.

	Coefficient estimate	p-value	95% confidence interval
Age.na	-0.263	0.00260	$(-0.434, -0.092)$
GenderNA	-0.329	0.000144	$(-0.499, -0.159)$
EmploymentCategoryNA	-0.407	5.54e-08	$(-0.554, -0.260)$
EmploymentCategoryPrivate Consultant/Self Employed	0.268	0.00820	$(0.069, 0.466)$
JSMtot	0.270	$<$ 2e-16	$(0.242, 0.299)$
USA.CAN	-0.270	1.53e-05	$(-0.392, -0.147)$
InChapter	1.583	$<$ 2e-16	$(1.509, 1.657)$
MEMTYPE_newB	1.654	0.00283	$(0.568, 2.740)$
MEMTYPE_newC	1.985	0.000223	$(0.931, 3.039)$
MEMTYPE_newD	1.914	0.000361	$(0.862, 2.966)$
P.SEC.BIOM	-0.562	0.00416	$(-0.946, -0.177)$
P.SEC.BIOP	-1.046	6.26e-06	$(-1.500, -0.592)$
P.SEC.CNSL	-0.472	0.0191	$(-0.867, -0.077)$
P.SEC.MHS	1.153	0.0187	$(0.192, 2.115)$
P.SEC.NPAR	0.729	0.0246	$(0.093, 1.365)$
P.SEC.SPES	-0.621	0.0491	$(1.240, -0.002)$
P.SEC.SRMS	-0.568	0.0131	$(-1.016, -0.119)$

Table 1: Coefficient summary with 95% confidence intervals for statistically significant ($\alpha = 0.05$) linear predictors of final GAM

2.3 Binary Classification Tree

For comparison, we also used a binary classification tree to model the response. Unlike linear and additive models, the final result of a tree-based model is a partitioning of the observations into distinct groups, with the groupings determined by ranges of values of the predictor variables. We first fit the tree model on all the predictors and set the complexity parameter `cp` to a low value (0.001). In order to prevent the tree from overfitting, we performed 10-fold cross-validation and used the 1-SE method to locate the largest `cp` value (the smallest tree) where the deviance statistic is less than the sum of the smallest deviance statistic and its associated standard error. As shown in Figure C.10, the optimal `cp` based on this method is 0.0037 and the final pruned model makes splits on `InChapter`, `Gender`, `JSMtot`, and `AgeJoinedASA`; these predictors were also found to be significant in the GAM.

2.4 Comparison of Models

Despite the inherent differences among tree-based models, GLMs, and GAMs, we wanted to compare the performance of our three models and decided to use classification accuracy on a held out test set. The results are summarized in Table 2. The test accuracies are are all quite similar, but the GAM slightly outperforms the GLM and classification tree with a classification accuracy of 0.7078. This provides additional evidence, along with the preference of the GAM over the GLM based on the likelihood ratio test conducted in Section 2.2, that the GAM should be used as our final model.

Model	Test accuracy
GLM	0.7047
GAM	0.7078
Classification tree	0.7001

Table 2: Classification accuracy on the test set for all models

3 Discussion

3.1 Interpretation of Final Model

Based on the summary of significant coefficient estimates shown in Table 1, our final GAM allows us to draw several important conclusions about the relationships between the predictors and the probability of ASA members belonging to at least one section. First, we notice that several of the coefficient estimates related to missing values are negative. In particular, the missing value indicator `Age.na` suggests that the presence of a missing value of `Age` is associated with a decrease of 0.263 in the probability of being a member of any section on the logit scale beyond the mean of the observed ages. Similarly, the estimates for the `GenderNA` and `EmploymentCategoryNA` factor levels indicate that not entering values for `Gender` and `EmploymentCategory` are associated with decreases of 0.329 and 0.407 in the probability of being a member of any section on the logit scale relative to being female and having an "academic" employment category, respectively. As hypothesized in Section 1.1 (Exploratory Data Analysis), these relationships may exist because ASA members who spend less time filling out their personal information are possibly less engaged with the ASA in general, and therefore less likely to belong to a section.

Next, several of the positive coefficient estimates also have important interpretations. Unsurprisingly, `InChapter` and `JSMtot` both have significant positive estimates which suggest that being a current chapter member is associated with a 1.58 increase in the probability of being a member of any section on the logit scale, and increasing the number of Joint Statistical Meetings attended by an ASA memeber by one is associated with a 0.270 increase on the logit scale, respectively. It also appears that being a private consultant or self employed is associated with a 0.268 increase on the logit scale, relative to being in an academic employment category. Further, coefficient estimates for the `MEMTYPE_new` factor levels are all large positive values, which indicate that having a membership type in groups B, C, or D is associated with an increase between 1.65 and 1.99 in the probability of being a member of any section (on the logit scale) relative to group A, which includes IFREP (Institutional Representative - Faculty), ILIFR (Life Membership - Retired), ISREP (Institutional Representative - Student). While it is unsurprising that retired members are less likely to be involved in sections, it is interesting to note that both student and faculty institutional representatives also appear to be less engaged than other membership types.

The coefficient estimates for section indicators provide insights into the relative engagement of various sections. In particular, being in the Biometrics and Biopharmaceutical sections at some point prior to March 2019 are associated with 0.562 and 1.05 decreases in the probability of being a section member (on the logit scale) in March 2020, respectively. This suggests that ASA members working in the biotechnology space may now be less engaged in sections than they used to be; we see similar effects for the Consulting, Physical and Engineering Sciences, and Survey Research Methods sections, which also have negative coefficient estimates. On the other hand, being a member of the Mental Health Statistics or Nonparametric sections before March 2019 is associated with 1.15 and 0.729 increases in the probability on the logit scale of being a current member of any section.

Finally, the smooths shown in Figure 2 provide a visual interpretation of the true relationships between the age variables and the response. More specifically, we see that between the ages of 20 and 60, ASA members are increasingly likely to belong to at least one section. Beyond the age of around 65, however, the probability of belonging to a section appears to decrease linearly. Given that many ASA members likely retire around this age, this trend agrees with our observation that having a "Life Membership - Retired" membership is also associated with a lower probability of belonging to a section. In addition, the smooth on the right of Figure 2 shows that members who joined the ASA at a younger age are more likely to belong to a section, which makes similar intuitive sense.

3.2 Recommendations

Based on the interpretations from our final GAM, we propose a set of recommendations that might help the ASA boost section engagement:

- First, there is strong evidence that increasing chapter membership and Joint Statistical meeting attendance would increase the number of people involved in sections. This could be done by hosting more

in-person meetings or marketing gatherings by advertising free food and networking opportunities.

- Second, we found that both student and faculty institutional representatives were less likely to belong to a section. Given the ASA's many connections to universities, it might be fairly easy to target these members and encourage them to join sections. Further, all the section predictors with negative coefficient estimates should be examined. Section leaders in the biotechnology space, in particular, should be evaluated and provided greater guidance about how to encourage and maintain engagement.
- Third, there is strong evidence that members who join the ASA at a younger age are more likely to be engaged in sections. Therefore, the ASA could target recent graduates who are just beginning their careers in statistics. Offering networking and mentorship opportunities at recruiting events, for example, is likely to entice more students to join.
- Fourth, it is important to note that our analysis is based on data collected before the COVID-19 pandemic. Due to the significant impact of the pandemic, it is possible that our conclusions about the relationships between various predictors and the response are less valid in a post COVID world. Therefore, our final recommendation is that new data be collected for an updated analysis in the next few years.

3.3 Evaluation

Our analysis yielded a final GAM-based approach for predicting whether an ASA member belonged to any section with a predictive accuracy of 70.78%. There were, however, some unexpected results and limitations to our work. First, we expected the three modeling approaches (GLM, GAM, and classification tree) to produce a greater diversity of test accuracies. As shown in Table 2, they were all within 1 percentage point of each other. Second, using a GAM made it difficult to interpret the interactions between `Age` and `EmploymentCategory` and between `Age` and `JSMtot`, both of which were statistically significant. A future study might consider investigating the non-linear interactions between these variables more closely. Finally, while we were limited by time constraints, other groups could examine an even greater range of preprocessing methods for this data set, especially for handling outliers, missing values, and categorical groupings.

4 Appendix

A Binary response GLMs tested

Model 1: `Age+AgeJoinedASA+Age.na+Gender`

Model 2: `Age+AgeJoinedASA+Age.na+Gender+EmploymentCategory+JSMtot+USA.CAN`

Model 3: `Age+AgeJoinedASA+Age.na+Gender+EmploymentCategory+JSMtot+USA.CAN+DontPublish+InChapter +MEMTYPE_new`

Model 4: `Age+AgeJoinedASA+Age.na+Gender+EmploymentCategory+JSMtot+USA.CAN+DontPublish+InChapter +MEMTYPE_new+P.SEC.BIOM+P.SEC.BIOP+P.SEC.EPI+P.SEC.HPSS+P.SEC.MDD+P.SEC.MHS`

Model 5: `Age+AgeJoinedASA+Age.na+Gender+EmploymentCategory+JSMtot+USA.CAN+DontPublish+InChapter +MEMTYPE_new+P.SEC.COMP+P.SEC.GRPH+P.SEC.NPAR+P.SEC.QP+P.SEC.SBSS+P.SEC.SI +P.SEC.SLDM+P.SEC.SRMS+P.SEC.SSPA`

Model 6: `Age+AgeJoinedASA+Age.na+Gender+EmploymentCategory+JSMtot+USA.CAN+DontPublish+InChapter +MEMTYPE_new+P.SEC.BE+P.SEC.CNSL+P.SEC.EDUC+P.SEC.ENVR+P.SEC.GOVT+P.SEC.MKTG +P.SEC.SBSS+P.SEC.SDNS+P.SEC.SIS+P.SEC.SOC+P.SEC.SPES`

Model 7: `Age+AgeJoinedASA+Age.na+Gender+EmploymentCategory+JSMtot+USA.CAN+DontPublish+InChapter +MEMTYPE_new+P.SEC.BE+P.SEC.BIOM+P.SEC.BIOP+P.SEC.CNSL+P.SEC.COMP+P.SEC.EDUC+P.SEC.ENVR +P.SEC.EPI+P.SEC.GOVT+P.SEC.GRPH+P.SEC.HPSS+P.SEC.MDD+P.SEC.MHS+P.SEC.MKTG+P.SEC.NPAR +P.SEC.QP+P.SEC.SBSS+P.SEC.SDNS+P.SEC.SGG+P.SEC.SI+P.SEC.SIS+P.SEC.SLDM+P.SEC.SOC +P.SEC.SPES+P.SEC.SRMS+P.SEC.SSPA+P.SEC.TSHS`

Model 8: `Age+AgeJoinedASA+Age.na+Gender+EmploymentCategory+JSMtot+USA.CAN+DontPublish+InChapter +MEMTYPE_new+P.SEC.BE+P.SEC.BIOM+P.SEC.BIOP+P.SEC.CNSL+P.SEC.COMP+P.SEC.EDUC+P.SEC.ENVR +P.SEC.EPI+P.SEC.GOVT+P.SEC.GRPH+P.SEC.HPSS+P.SEC.MDD+P.SEC.MHS+P.SEC.MKTG+P.SEC.NPAR +P.SEC.QP+P.SEC.SBSS+P.SEC.SDNS+P.SEC.SGG+P.SEC.SI+P.SEC.SIS+P.SEC.SLDM+P.SEC.SOC +P.SEC.SPES+P.SEC.SRMS+P.SEC.SSPA+P.SEC.TSHS+Age:EmploymentCategory+Age:JSMtot`

B Deviance values for GLMs tested

Model	Deviance	Linear Coefficients
1	17344	5
2	16686	13
3	15111	18
4	15071	26
5	15094	27
6	15090	29
7	15039	45
8	15003	52

Table 3: Deviance values for binary response GLMs considered

C Plots and graphs

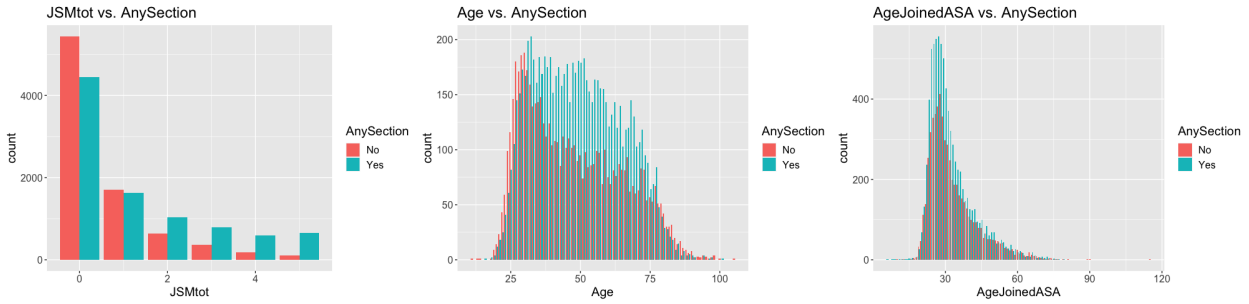

Figure C.1: Bivariate plots of numerical predictors versus response variable

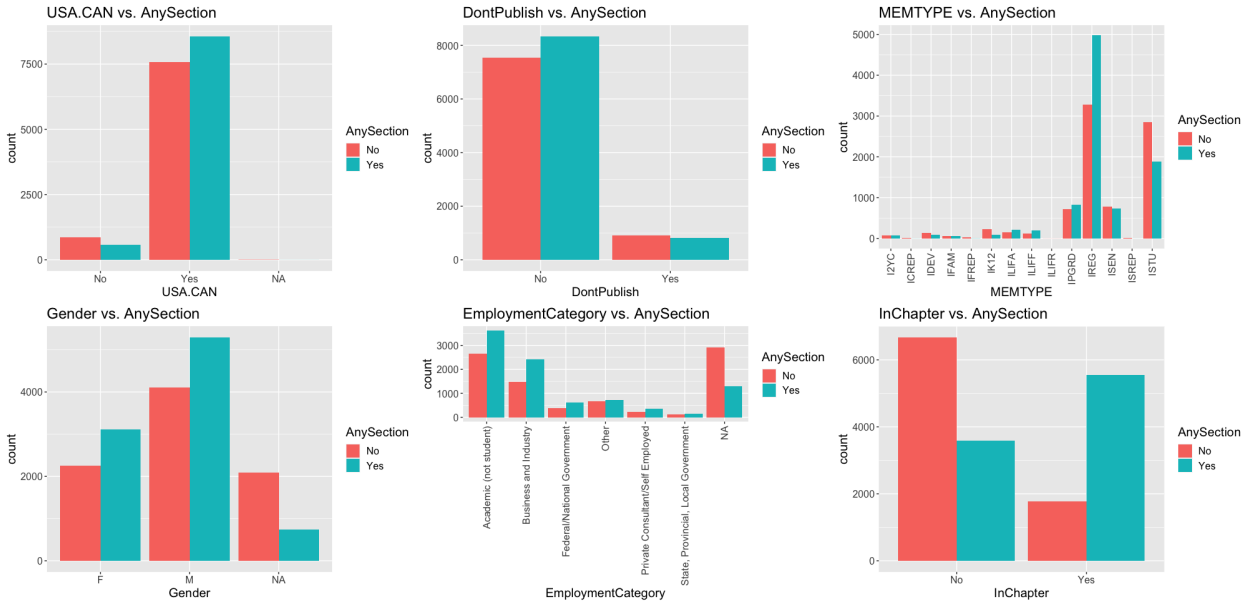

Figure C.2: Bivariate plots of categorical predictors versus response variable

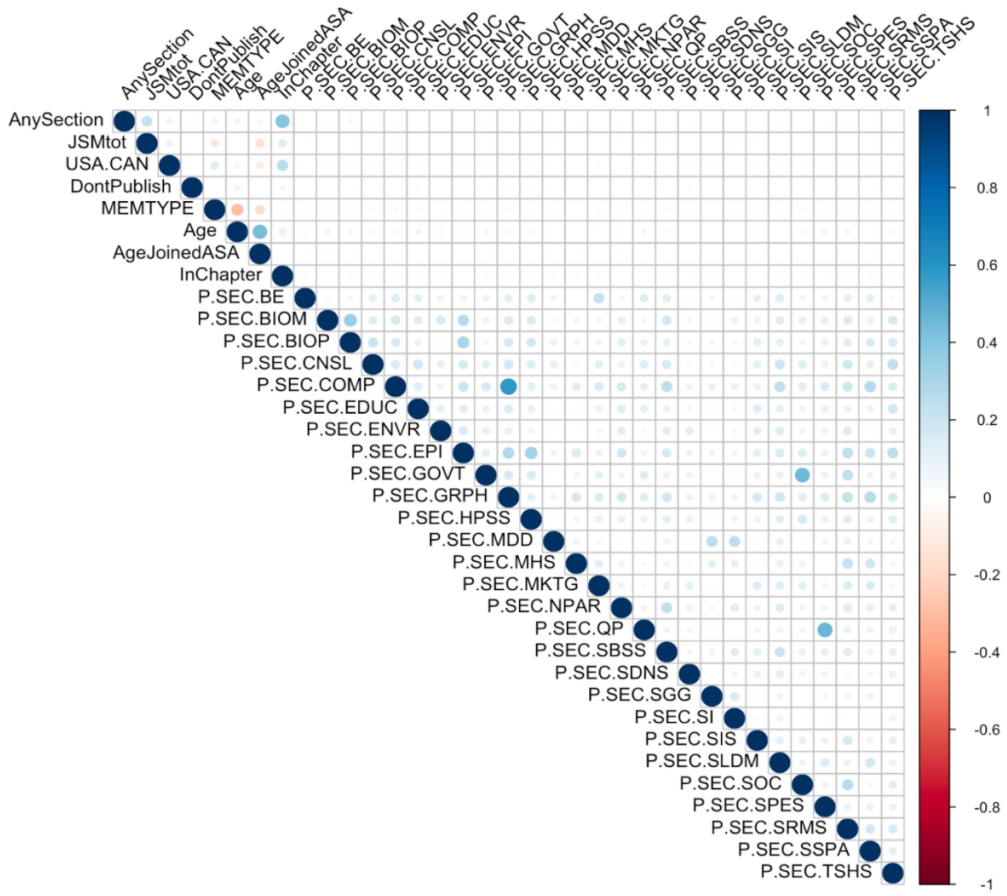

Figure C.3: Correlation matrix of all variables

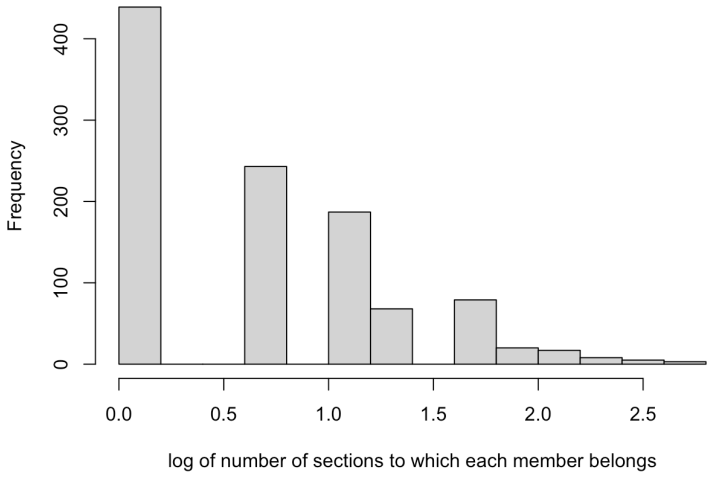

Figure C.4: Histogram of section membership counts (log scale)

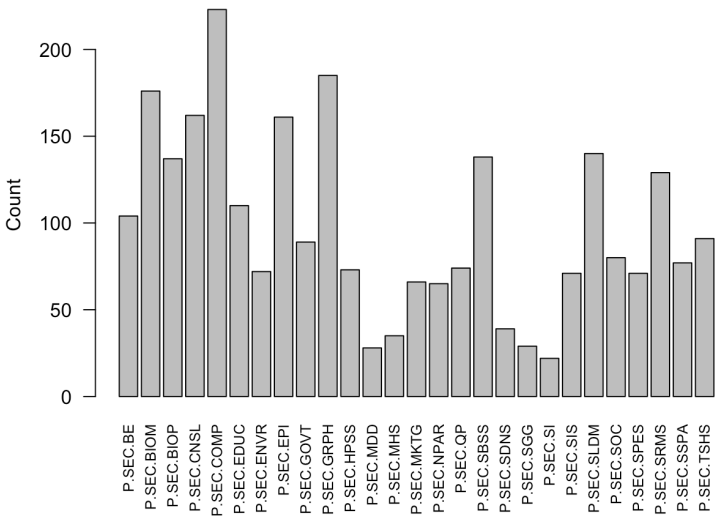

Figure C.5: Bar plot of number of members in each section

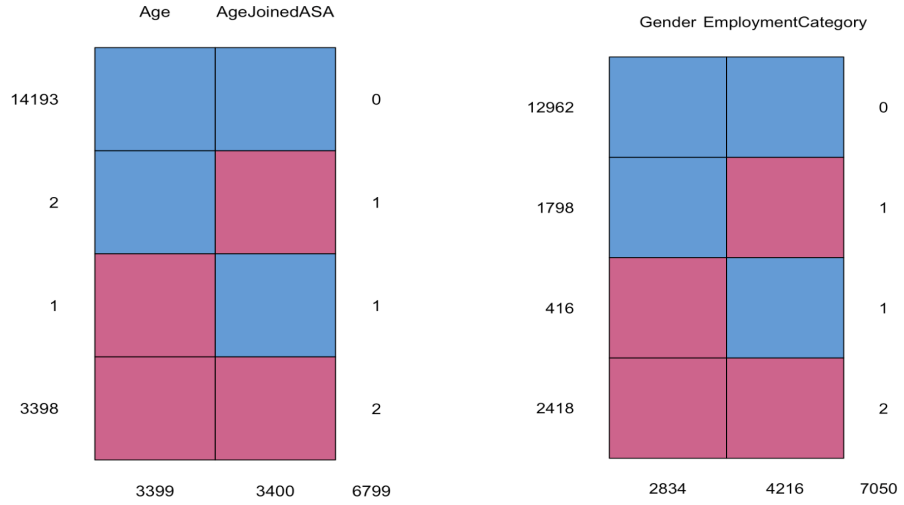

Figure C.6: Correlation matrix of missing values

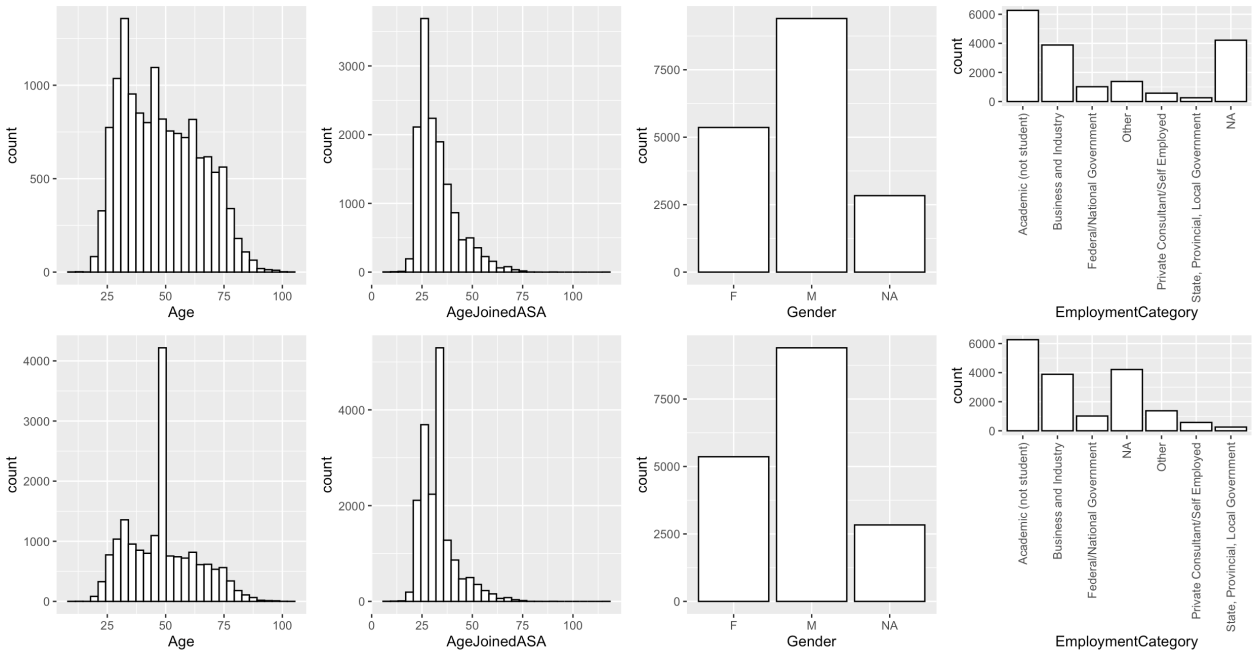

Figure C.7: Distribution of columns with missing values before (top) and after (bottom) GLM-based imputation

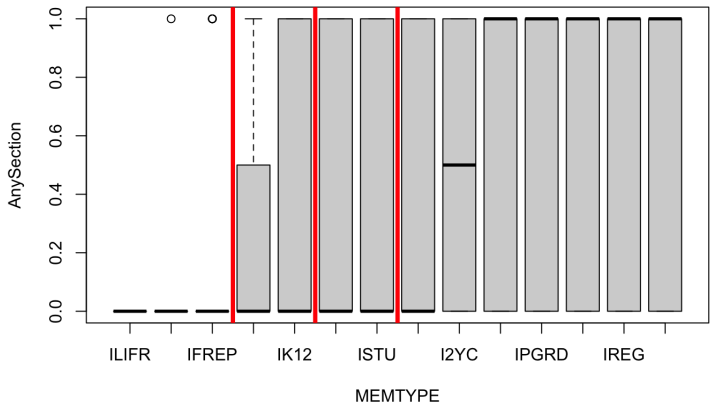

Figure C.8: Grouped membership levels

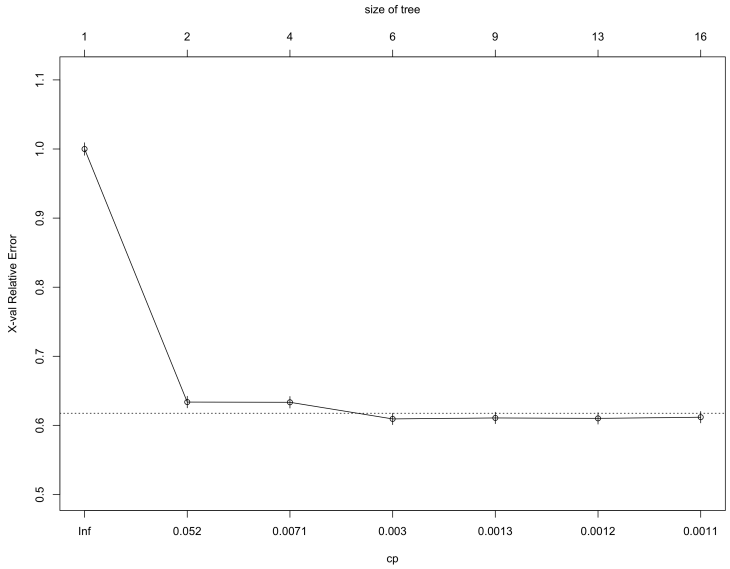

Figure C.9: Deviance statistic and standard errors for different `cp` values

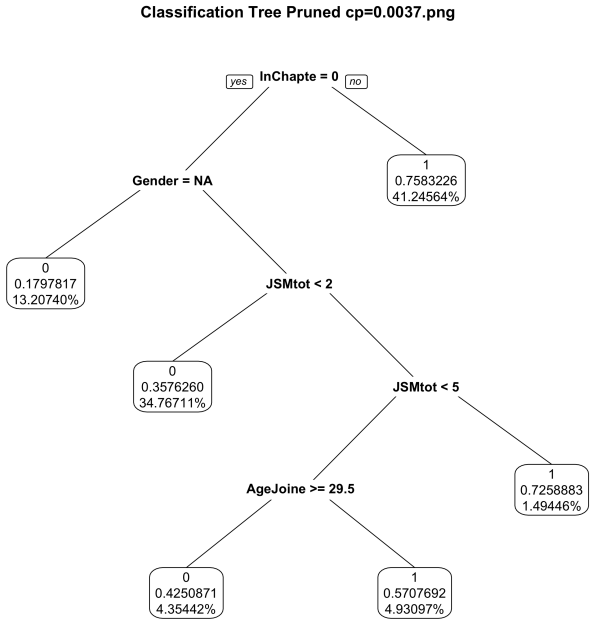

Figure C.10: Pruned classification tree model