2018 Qualifying Exam

Question 1

Part A.

Table 1 shows counts and percentages of subjects with each disease status, by phase of study in which they were enrolled. The table shows that only subjects with mild cognitive impairment (MCI) were enrolled in phase 2. Generally, the number of patients of each disease type is disproportionate across phases.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  |  |  | | --- | --- | --- | --- | --- | | |  | | --- | | **Frequency** | | **Percent** | | **Row Pct** | | **Col Pct** | | | | **Table 1: Phase by Disease** | | | | | | --- | --- | --- | --- | --- | | **Phase** | **Disease** | | | | | **AD** | **MCI** | **NL** | **Total** | | **1** | |  | | --- | | 178 | | 15.95 | | 23.70 | | 86.83 | | |  | | --- | | 368 | | 32.97 | | 49.00 | | 61.03 | | |  | | --- | | 205 | | 18.37 | | 27.30 | | 66.56 | | |  | | --- | | 751 | | 67.29 | |  | |  | | | **2** | |  | | --- | | 0 | | 0.00 | | 0.00 | | 0.00 | | |  | | --- | | 98 | | 8.78 | | 100.00 | | 16.25 | | |  | | --- | | 0 | | 0.00 | | 0.00 | | 0.00 | | |  | | --- | | 98 | | 8.78 | |  | |  | | | **3** | |  | | --- | | 27 | | 2.42 | | 10.11 | | 13.17 | | |  | | --- | | 137 | | 12.28 | | 51.31 | | 22.72 | | |  | | --- | | 103 | | 9.23 | | 38.58 | | 33.44 | | |  | | --- | | 267 | | 23.92 | |  | |  | | | **Total** | |  | | --- | | 205 | | 18.37 | | |  | | --- | | 603 | | 54.03 | | |  | | --- | | 308 | | 27.60 | | |  | | --- | | 1116 | | 100.00 | | |

Figure 1 shows a plot of a principal components analysis (PCA) of all ROI biomarkers. Phase 1 subjects are shown in red, phase 2 subjects are shown in green and phase 3 subjects are shown in blue. The values for the first two principal components for each subject, calculated using all ROI biomarkers, are plotted against each other to represent similarities between subjects and show any clusters of subjects with similar traits. The principal components for each subject represent the variability in ROI biomarker values. Principal component 1 explains 31.24% of the variability in response, and principal component 2 explains 6.52% of the variability in response. Though there is some overlap, it is clear that there are batch effects, and subjects in phase 1 (represented more in upper left of the plot) tend to have some graphical separation from subjects in phase 3 (represented more in lower right of the plot).

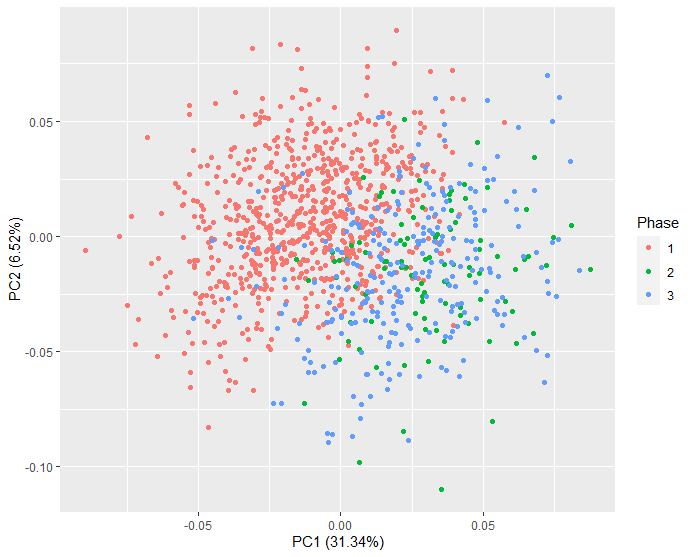


Figure 1: Principal components plot

Part B.

We would like to assess whether there is a strong association between SNPs and disease status. For each SNP marker we fit a separate proportional odds model on the ordinal disease status {NL < MCI < AD}, while adjusting for the same set of covariates in each model: Age, Gender, gPCs and Phase.

The proportional odds model is specified as such:

Where is the probability of AD, is the probability of MCI, and is the probability of NL. The reference level for gender is male, the reference for phase is 3, and the reference for SNP category is 2. The assumptions of the proportional odds model are as follows:

1. The two logits have different intercepts ≠ .
2. Proportional Odds Assumption: The regression coefficients ( in the model) are equal for both logits.

Table 2 shows the odds ratios and 95% confidence intervals for SNPs 1-10. The odds ratio represents the odds of a less favorable condition, for subjects where SNP=1 compared to subjects where SNP=2, holding other model covariates constant.

Table 2: ORs and 95% CI for SNP models

|  |  |  |  |
| --- | --- | --- | --- |
| SNP | OR | Lower | Upper |
| SNP1 | 0.3611 | 0.2812 | 0.4636 |
| SNP2 | 0.3561 | 0.2236 | 0.5671 |
| SNP3 | 0.3916 | 0.3074 | 0.4988 |
| SNP4 | 0.3309 | 0.2177 | 0.503 |
| SNP5 | 0.3657 | 0.2848 | 0.4697 |
| SNP6 | 0.3584 | 0.2228 | 0.5765 |
| SNP7 | 0.3657 | 0.2848 | 0.4697 |
| SNP8 | 0.3584 | 0.2228 | 0.5765 |
| SNP9 | 0.3705 | 0.2884 | 0.476 |
| SNP10 | 0.352 | 0.2181 | 0.5683 |

It should be noted that the test for the proportional odds assumption for all of these models had a p value of <0.0001. It is not reasonable to assume that regression parameters in the model(s) are equal across the two logits.

Part C.

We would like to carry out a goodness of fit analysis for these models using different link functions, the probit and complementary log-log link. Table 3 shows the log likelihood values for each of the 3 links for all 10 SNP models. The models using the complementary log-log link appear to consistently have the smallest deviance, which suggests they may have the best fit.

Table 3: Model Fit by Deviance

| **SNP** | **Logit** | **Probit** | **Complementary Log-Log** |
| --- | --- | --- | --- |
| 1 | 2121.5804 | 2122.7046 | 2109.9880 |
| 2 | 2169.0491 | 2168.1553 | 2148.6082 |
| 3 | 2128.2129 | 2128.3745 | 2115.2685 |
| 4 | 2161.0692 | 2159.9576 | 2140.5587 |
| 5 | 2123.4086 | 2124.6193 | 2110.9224 |
| 6 | 2170.0766 | 2169.1992 | 2150.3274 |
| 7 | 2123.4086 | 2124.6193 | 2110.9224 |
| 8 | 2170.0766 | 2169.1992 | 2150.3274 |
| 9 | 2125.3811 | 2126.4925 | 2111.5642 |
| 10 | 2169.7118 | 2168.8669 | 2150.2405 |

Part D.

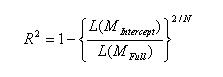
We now assess goodness of fit using pseudo R squared statistics. Table 4 shows the Cox Snell and Nagelkerke pseudo R squared statistics for all 10 models using the logit link.

Table 4: Pseudo R squared statistics for models with logit link

|  |  |  |
| --- | --- | --- |
| **SNP** | **Cox Snell psuedo R-squared** | **Nagelkerke psuedo R-squared** |
| 1 | 0.092732 | 0.107273 |
| 2 | 0.053309 | 0.061668 |
| 3 | 0.087323 | 0.101017 |
| 4 | 0.060054 | 0.069471 |
| 5 | 0.091244 | 0.105552 |
| 6 | 0.052437 | 0.060659 |
| 7 | 0.091244 | 0.105552 |
| 8 | 0.052437 | 0.060659 |
| 9 | 0.089636 | 0.103693 |
| 10 | 0.052746 | 0.061018 |

The key element of the Cox Snell pseudo R squared statistic represents a ratio of the likelihood of the intercept-only model against the full model. A smaller ratio indicates a greater improvement of model fit over the intercept-only model. Thus, a larger value for the statistic indicates better model fit. In our 10 SNP models this statistic is consistently < 0.1, indicating the full model does not improve much over the intercept-only model.

Equation 1: Cox Snell pseudo R squared



The Nagelkerke pseudo R squared statistic is a normalized Cox Snell statistic such that the range of possible values is 0 to 1, where 1 indicates the best model fit, which would predict the data exactly. Again we see this statistic indicates a poor model fit.

Part E.

We would like to perform model-selection to determine the best fitting model from all of the available variables, to predict disease status, modeled as a multi-categorical outcome. Using the LASSO method with a criterion of minimizing AIC, we obtain a model with the following covariates:

Intercept SNP1 SNP2 SNP3 SNP4 SNP5 SNP6 SNP7 SNP8 SNP9 Age Gender Phase

It is interesting to note that none of the ROI biomarkers were selected as predictive variables, nor were any of the genetic principle components. Note: NL was used as the reference disease level in the model. A slightly different model (including genetic PC5) is selected if AD is used as the reference.

Part F.

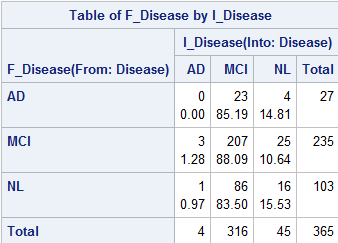
i. We would like to use the phase 1 data to predict the disease status of subjects in phases 2 and 3. We would like to find the optimal predicted disease status which minimizes the number of incorrect classifications. The objective function of the prediction model is the following:

D\_N(hat) = argmax{X’\_N \* alpha\_k (hat)}

Where X\_N is the data for a new subject and alpha\_k(hat) is the vector of estimated regression coefficients for disease status k obtained from fitting the model using phase 1 subjects.

Table 5: Actual vs Predicted Disease Status (Fi)

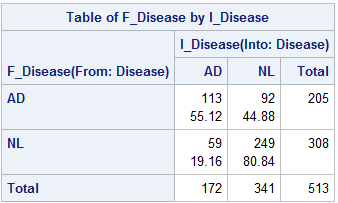
Table 5 shows the crosstabulation of actual (F\_Disease) vs predicted (I\_Disease) after using the variables selected from Part E, the regression parameters calculated using phase 1 subjects, and the data from phase 2 and 3 subjects. It appears the subjects with AD or NL in phase 2 and 3 subjects are generally misclassified. Subjects with MCI were generally correctly classified.



ii. We now exclude subjects with MCI disease status from the data set and use a logistic regression model for prediction. The MCI disease status is strongly associated with phase 2 (phase 2 only used subjects with MCI disease status) that we may get better predictive results for NL and AD if we exclude MCI data.

Table 6: Actual vs Predicted Disease Status (Fii)

Table 6 shows the crosstabulation of actual (F\_Disease) vs predicted (I\_Disease) after using the variables selected from Part E, the regression parameters calculated using phase 1 subjects, and the data from phase 2 and 3 subjects. Approximately 55% of subjects with AD and 81% of subjects with NL were correctly classified using this logistic model, excluding subjects with MCI.



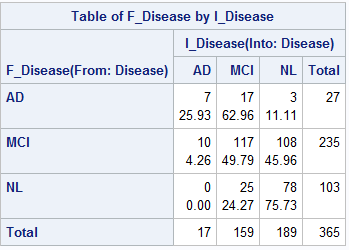
Part G.

Next we attempt to use a Random Forest method for model selection instead of LASSO. Using all of the variables in the data set and all subjects, the top 7 variables in order of importance are:

Age, ROI99, ROI30, ROI41, ROI44, SNP9, SNP4

In an attempt to create a relatively parsimonious model for the 1116 subjects in the data set, the top 7 variables were chosen, such that the top 2 SNPs were included. We find that after reducing the data set by selecting for phase, some of the SNPs are highly colinear.

Table 7: Actual vs Predicted Disease Status (G)

Table 7 shows the crosstabulation of actual (F\_Disease) vs predicted (I\_Disease) after using the variables selected from the random forest model, the regression parameters calculated using phase 1 subjects, and the data from phase 2 and 3 subjects. This model correctly classifies a greater percentage of AD (26% vs 0%) and NL (76% vs 16%) subjects, at the expense of higher accuracy of MCI (50% vs 88%) classification, as compared to the LASSO model.

Part H.

In this study SNP and ROI biomarker data is collected on patients classified as having Alzheimer’s disease, mild cognitive impairment or normal status. We attempt to find predictors of disease status using these markers, while controlling for gender, age, genetic principle components, and phase. We identified batch effects from the ROI biomarker data related to phase. Because only patients classified as having MCI were included in phase 2, simply using phase as an adjustment variable in our regression models may not completely account for the effects of phase when trying to predict if patients would be classified as AD or NL. To have a more balanced design, it would have been advantageous to recruit equal numbers of patients with each disease class across phase, though this is generally not possible in many observational studies.

When using machine learning techniques such as LASSO and random forest, we find that each method selects a different set of covariates and prediction accuracy varies. The LASSO model was able to classify MCI patients more accurately, whereas the random forest model appeared to classify AD and NL patients more accurately, at the expense of a higher misclassification rate of MCI patients.

Ultimately, the choice of predictive model for clinical applications would have several considerations. First, what is the trade-off between increased prediction accuracy and clinical cost to collect the data needed for these variables? Second, is it more important to classify patients at the extremes (AD vs NL) rather than MCI?

It is also important to note the limitations of using the proportional odds model when examining SNP data. For all of the models, the test of the proportional odds assumption indicated that there were widely varying odds ratios for all SNPs between levels of disease. When running a partial proportional odds model (data not shown) the two odds ratios for the two logits were very different. If examining effect of single nucleotide polymorphisms on disease status were the primary focus of the study, more exploratory analysis would be warranted. It is also likely that classification as an ordinal variable is not a useful way to model this data.