Andrew J. Spieker, PhD
BIOS 6312 - Modern Regression Analysis
Spring 2020
Homework #6 Key

Consider the MRI study (`mri.csv`). We seek to develop a predictive model for global brain atrophy, defined as an atrophy score exceeding 35. We'll explore and compare a few ways to do this.

1. Let $Y = 1(\text{atrophy} > 35)$ denote global brain atrophy. Consider the following models for Y:

   (I) A logistic model with only an intercept.

   (II) A logistic model including gender, race category, stroke category, and a restricted cubic spline on age (with knots at 70, 75, and 80 years).

   (III) A logistic model including terms for age, gender, race category, weight, coronary heart disease category, stroke category, low-density lipoprotein, blood albumin, blood creatinine, platelets, and FEV, including a LASSO penalty selected by cross-validation.

   (a) One way to characterize prediction error under binary outcomes is $E = \frac{1}{N}\sum_i (y_i - \widehat{p}_i(\boldsymbol{x}_i))^2$, where $\widehat{p}_i(\boldsymbol{x}_i) = \widehat{\mathrm{P}}(Y_i = 1|\mathbf{X}_i = \boldsymbol{x}_i)$. With this in mind, split the data into random halves, a "training" and a "test" set, using a seed of $s = 6312$. In turn, fit Models (I), (II), and (III) to the training data, employing five-fold cross-validation for Model (III) within the training set (with a seed of $s = 2020$) to select an appropriate tuning parameter. Report the training and test prediction errors for each model.

   Ans: The results are reported in the table below:

Method	Training Error	Test Error
Model (I)	0.2500	0.2502
Model (II)	0.2179	0.2260
Model (III)	0.2192	0.2261

   (b) Briefly comment on your findings—do they align with your expectations?

   Ans: There are a few things I would have expected *a priori*. First, I would expect the test error to generally be no smaller than the training error. Second, I would expect the prediction errors in Models (II) and (III) to be lower than those in Model (I). Finally, I would expect that the difference in training and test error for Models (II) and (III) to generally be no smaller than that of Model (I)—the idea being that there's no reason to believe that a null model with just an intercept should have dramatically greater performance on the training set from which it was fit than on a test set. The patterns seen in the table above confirm my expectations.

   I was generally anticipating that people would have made note of the first and second points listed above.

2. Now, we will tackle the problem from another angle (namely, by directly modeling the continuous outcome, Y = atrophy, and *then* dichotomizing). Keeping in mind the goal of predicting global brain atrophy, define the prediction error to be $E = \frac{1}{N}\sum_i |1(y_i > 35) - 1(\widehat{Y}_i(\boldsymbol{x}_i) > 35)|$, where $\widehat{Y}_i = \widehat{\mathbf{E}}[Y_i|\mathbf{X}_i = \boldsymbol{x}_i]$. Consider the following models:

   (I) A linear model with only an intercept.

   (II) A linear model including gender, race category, stroke category, and a restricted cubic spline on age (with knots at 70, 75, and 80 years).

   (III) A linear model including terms for age, gender, race category, weight, coronary heart disease category, stroke category, low-density lipoprotein, blood albumin, blood creatinine, platelets, and FEV, including a LASSO penalty selected by cross-validation.

   (a) State the number of degrees of freedom used by each of these three models.

   Ans: The model degrees of freedom are 1, 9, and 5, respectively.

   (b) Based on the training/test sets you created in Problem 1, and using the same seeds, fit Models (I)-(III). Report the training and test prediction errors for each model. How do these results compare to those of Problem 1?

   Ans: The results are reported in the table below:

Method	Training Error	Test Error
Model (I)	0.4946	0.5150
Model (II)	0.3505	0.3597
Model (III)	0.3533	0.3842

We see generally similar patterns, although in this example, we note that the training and test error are quite close in Model (II) and notably further apart for Model (III), possibly just owing to random variation. It's no surprise that the spline and LASSO methods tend to have better performance in these problems. Recall that they both attempt to optimize predictions via the bias-variance trade-off (the former choosing to reduce the bias at a cost of a higher variance, and the latter choosing to reduce the variability at a cost of bias). I would expect that spline-based regression might perform better than penalized regression in settings where there were a few covariates whose relationship with the mean outcome were highly nonlinear, and penalized regression might perform better than spline-based regression in settings where there are a large number of covariates relative to the sample size at hand.

3. Compare the approaches of Problems (1) and (2) and their relative advantages. Things to consider include the choice of dichotomizing "before" or "after," and prediction error definition.

Ans: The clinical question at hand pertains to a dichotomous definition. In Problem (1), we dichotomize our outcome prior to modeling and then formulate a definition for prediction error (of which there are many) suitable for binary outcome regression. In Problem (2), we model the continuous outcome and then dichotomize the predictions, formulating a suitable prediction error definition. In the first case, we lose information right away by failing to use the continuous outcome, but in the second case, the definition for prediction error is much more rigid (either zero or one for each subject).