MS&E 226 Mini-Project

Hershel Mehta & Sara Altman

Regression

The problem: We consider a clinically relevant problem: predicting how many days after a rehabilitation program someone remains sober.

Our outcome measure, `obstime`: As our continuous outcome we used the variable `obstime`, which encodes how many days after the rehabilitation program we observed sobriety in each patient. In our models, we encounted many challenges working with this covariate, since

1) patients are often lost to follow-up (meaning they don't answer our calls to record their latest sobriety data), making it difficult to confirm whether or not they've relapsed or not.

2) `obstime` by its nature is an "observed" time, meaning it's not exactly the number days of sober, but the number of days when we confirmed that the patients were sober. This will be even more of a challenge later on, when we deal with interpretation.

Here, we describe how we've dealt with the challenges of modeling `obstime`.

Modeling Approach: Since we were working with just a few subjects, we wanted to start with a simple model using the most salient features, look at our resulting residuals to see if we missed any big trend, and then add more features as needed.

Our evaluation metric was RMSE. RMSE is meaningful in this context–it represents how many days of sobriety on average our prediction would deviate from the true number of days of sobriety. Currently, there are no standard measures used to predict time to relapse, so we used the RMSE for a baseline model that predicts the mean of `obstime`. This baseline model has RMSE = 684.62, so its predictions would be off by 2 years on average. To be clinically relevant, we wanted a model with RMSE of less than 30 days.

When comparing models to each other, we used k-fold cross-validation with $k = 5$. We used a smaller k here since we didn't care as much about bias when comparing between models, and wanted to reduce variance. We also wanted to check for overfitting. When testing our final model, we used LOOCV, however. This is because we wanted an unbiased estimate of the test error, and it was computationally feasible to do LOOCV with our data set size.

Feature Selection: We had reason to believe a-priori that the brain region called the Nucleus Accumbens (NAcc; essentially, a brain region that is marker of dopamine response and anticipating rewards) would be the most pertinent of the brain regions for predicting `obstime`. There's two main approaches to extract information from a brain region: 1) extract the raw time series for each type of trial (for example, visualizing how activity changes at each time interval for all "drugs" trials; see "Time Series" plot in EDA below); and 2) model how much activation in the whole trial is above baseline for each type of trial and extracting beta coefficients (for example, coefficient representing how much "drugs" trials are above baseline; see "Density Plots of Betas" plot in EDA below)

You could theoretically include both approaches. However, it is considered best practice only use one in the neuroscience field. Therefore, we needed to figure out which approach best predicted `obstime`. To address that problem (and to determine the most predictive features generally), we used the Lasso.

From our Lasso results, we found that the most salient features for predicting `obstime` were: 1) `relIn6Mos` (an indicator variable corresponding to whether someone relapsed in 6 months after the program), and 2) `naccR_drugs_beta` (a continuous variable corresponding to the betas extracted from the right side of the NAcc, called the RNAcc). This is interesting because we included all of our covariates (including different brain regions using both approaches mentioned above), and the model selected the betas from the NAcc, confirming our a priori hypothesis.

We used those two variables as our starting point for our original model of `obstime`.

Our original model of `obstime`:

Here, our best model is a linear model with an interaction of the two features we found above. So Let $Y =$ `obstime`, $X_1 =$ `naccR_drugs_beta`, and $X_2 =$ `relIn6Mos`.

We found that our average RMSE with k-fold cross-validation where $k = 37$ was 73.54. Although that is better than our baseline model, it's still not quite clinically useful to say someone is going to relapse within a 75-day window.

To diagnose this issue, we carefully looked at our residuals and noticed that the main deviations came from non-relapsers, which makes sense because of the way `obstime` is encoded. If you assume that there is a real linear relationship between the measure of naccR response and the amount of time someone is sober, then we should expect our model to perform worse on non-relapsers then relapsers. This is because `obstime` doensn't actually encode the total amount of time a non-relapser is sober. They will likely to continue to be sober after the last time the researcher checked in with them.

We decided to only use regression in the context of our patient population who relapsed because we thought that would be the most clinically useful. A nice pipeline would be to assess the likelihood of someone relapsing or not within 6 months (see our classification section), and then use a regression built for potential relapsers to determine how many days they'll remain sober to determine whether they need to stay longer.

Our final model of `obstime` (note: constrained to only relapsers):

Here, our best model was a simple linear model where $Y =$ `obstime` and $X_1 =$ `naccR_drugs_beta`.

We found that our average RMSE with k-fold cross-validation where $k = 15$ was 34.83. Note that, here, our baseline model has an RMSE of 137.5.

These results were very exciting! Not only did we improve our RMSE from the baseline, but these results could also be used clinically to make a prediction that someone is going to relapse which is off by around 30 days (i.e., month). However, we have to be cautious to note that these results are drawn from only 15 subjects, so we would need to replicate in order to really see how clinically relevant these results are.

Conclusions, Learnings, and Thoughts on Bias-Variance:

We can predict days of sobriety within an error that is clinically relevant. We believe these results can help clinicians understand how long patients will stay sober.

In the process of learning how to deal with `obstime`, we looked into Survival Analyses. These models are built to deal with variables like `obstime`, where some observations may be lost to follow-up. We will try applyling these analyses to our data in the future.

Because we had so few subjects to work with after we filtered our dataset to only the relapsers, we had to consider how testing our model repeatedly might introduce bias into the modeling process. For that reason, I think our final RMSE may be an underestimate of the true error.

Classification

The problem:

We predict if a person will relapse in 6 months or not. Note that this is distinct from the problem of predicting whether or not someone will relapse at all, since some people relapse after 6 months. We chose six months in part because our data shows two groups of relapsers: those who relapse before 6 months and those who relapse after 6 months. Those who do not relapse are counted as not relapsing in 6 months, since they do not relapse at all.

This is also useful because 6 months gives an indication for a practioner if someone is likely to relapse soon, and so may require extra treatment now, or is likely to relapse far later or not at all, in which case the course of action required may be different.

Modeling approach:

Interpretability is important here. We wanted a model that was easily interpretable, as we're curious about the neuroscience behind this prediction problem.

The baseline to which to compare our model was the performance of a naive classifier that just classifies everyone as the most likely category. In our data set most people (25/39) do not relapse in 6 months. Therefore, if we predicted that no one would relapse in 6 months, our error rate (using 0-1 loss) would be $14/39 \approx 36\%$. In this case, every misclassified case would be a false negative, and so the false negative rate is also 36%.

For this problem, a false negative (i.e., predicting someone will not relapse in 6 months when they actually do) is worse than a false positive (i.e., predicting someone will relapse in 6 months when they actually don't). A false negative might lead to someone missing out on necessary immediate treatment. We decided that would be worse that someone getting immediate treatment he/she may not have needed. Therefore, we wanted to keep the number of false negatives low, even if it affected the overall estimated test error.

Best model:

Our best performing model was a logistic regression model with only one covariate: `nacc_neutral_beta`. When fit on all of our data, the intercept is -.387 and the coefficient on `nacc_neutral_beta` is 21.074.

When selecting a model, we only considered models with covariates that had no NA's. This is because we did not want to reduce our data set any further, as well as because we found that the features with NA's did not appear strongly correlated with relapse within 6 months. None of the fMRI covariates had any NA's, so we only ended up removing behavioral/demographic covariates from consideration.

To evaluate our models, we conducted 10-fold cross-validation and compared the cross-validated test error. Our objective was always 0-1 loss. We ran Lasso logistic regression because we wanted to reduce the number of covariates, as well as to see which covariates were important predictors of relapse within 6 months. We also conducted k-neartest neighbors with various values of k. We also fit logistic regression models with varying combinations of covariates (including the covariates selecting by the Lasso). However, none of our models had as low of a cross-validated test error a logistic regression model with `nacc_neutral_beta` as the sole covariate.

When using a threshold of .45, the cross-validated test error for this model is .177. The false negative rate is .123. This is therefore an improvement over the specified baseline, in terms of total misclassification rate and false negative rate. We used LOOVCV to compute this test error, after using 10-fold cross validation to compare across models. This is because when selecting a model we wanted to reduce variance, but didn't care much about bias. However, when estimating the test error we cared more about reducing bias. Our data set is also very small, and so LOOCV was feasible.

Our model is very simple. It therefore will have higher variance than a more complex model, and lower bias. This isn't necessarily ideal, since we don't really want a model that is sensitive to small changes in the data. However, for our other models to have an estimated test error near the one of our simple model, we had to include both beta and signal coefficients, which sacrifices interpretability and is frowned upon in the neuroscience field.

Our model uses a threshold of .45–we classified those with predicted responses greater than .45 as relapsing in 6 months. We thought a threshold of .5 wouldn't be ideal in this case, because we were wanted to reduce false negative rate, even if it meant increasing total error rate slightly. Therefore, we investigated the effect of changing the threshold on the total error and false negative rates. .45 minimized both total error and false negative rates.

This test error is likely an underestimate of the true test error because, as with our regression model, we looked at the data as a whole before deciding which covariates to use. For an unbiased result, we should have selected a model without the cross-validation holdout data included. However, we think the bias is likely to be small, given that our model is so simple and there are theoretical reasons why NAcc beta is predictive of relapsing in 6 months.