Part 3

Sara Altman and Hershel Mehta

12/5/2017

Prediction on the test set

We didn't have enough data for a test set, so we used all our data in Part 2. We used leave-one-out cross-validation to get an estimate of the test error. The results of this process are reported here.

Classification

CV error rate	CV false negative rate	CV false positive rate
0.24	0.08	0.16

Our estimated test error is .24.

Regression

RMSE_CV
34.83

Our estimated test RMSE is 34.8.

Inference

(a)

Logistic regression model fitted on all data

term	estimate	std.error	statistic	p.value
(Intercept)	-0.1308772	0.4246112	-0.3082284	0.7579086
nacc_neutral_beta	25.5717862	8.3958789	3.0457545	0.0023210

The coefficient on nacc_neutral_beta is significant at the $\alpha = .01$ level. The p-value is .002. This means that, if the null hypothesis is true, the probability of observing a Wald statistic as extreme as 3.046 is .002. Here, the null hypothesis is that the coefficient on `nacc_neutral_beta` is 0.

We only have one covariate. If the coefficient on this covariate really is 0, then the chance it will be significant at the $\alpha = .05$ level is 5% (and is 1% at the $\alpha = .01$ level). Therefore, the chance is low that this one covariate would be significant if the null were true. In contrast, if we had 100 covariates, we would expect 5 of them to be significant at the $\alpha = .05$ level and so we might not believe our results if it turned out that 5 of our

coefficient were significant.

Therefore, we think we can believe our results that the coefficient on `nacc_neutral_beta` is significant. Our p-value may be favorably biased, however, for reasons discussed in part (e).

(b)

Proportion significant (alpha = .05)	Proportion significant (alpha = .01)	Proportion significant (alpha = .001)
1	1	0

We didn't have enough data for a test set and used cross-validation in the prediction part of the project. So for this part, we again used cross-validation. We created 10 folds and fit our model on the training data associated with each fold. The table above gives the proportion of folds where the coefficient on `nacc_neutral_beta` was significant, for three α levels.

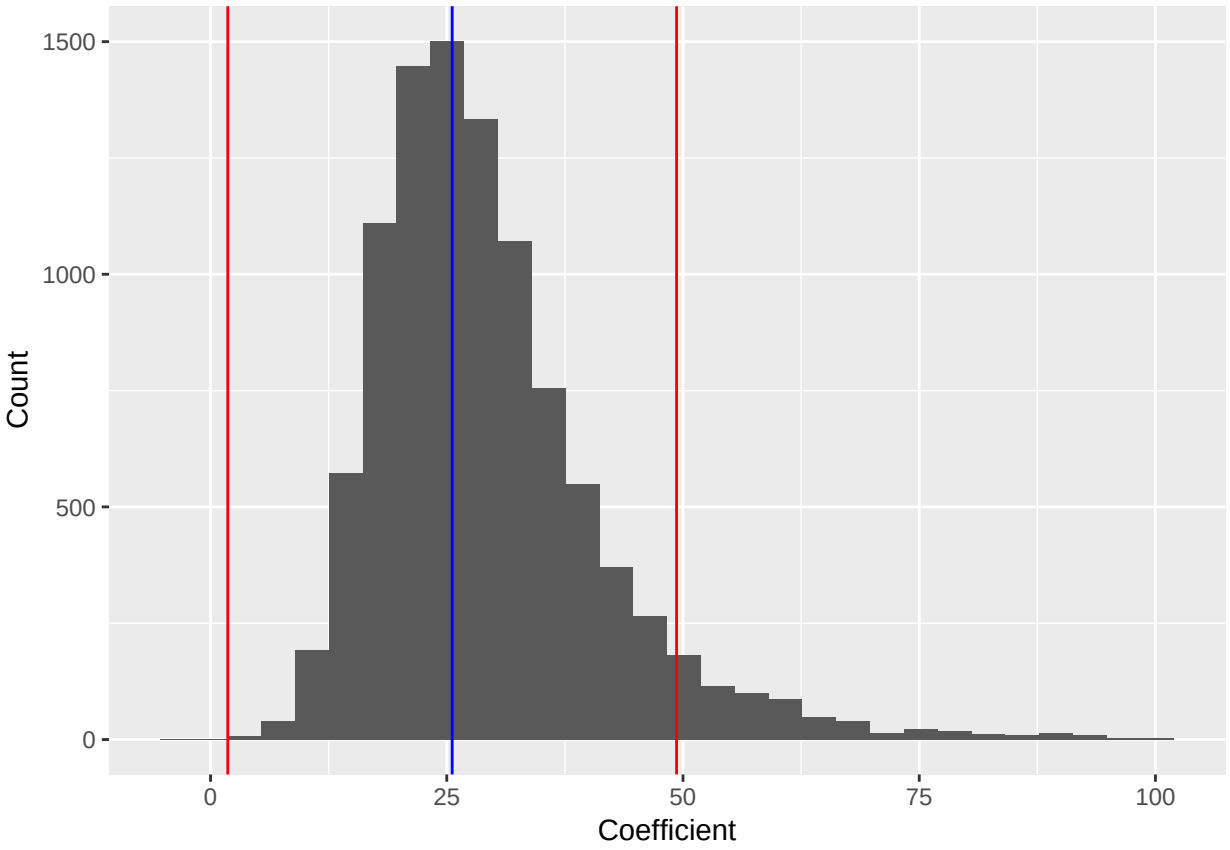
The coefficient on `nacc_neutral_beta` is significant in 100% of the folds at the $\alpha = .05$ level and at and is significant 90% of the time at the $\alpha = .01$ level.

(c)

Our original 95% confidence interval for our `nacc_neutral_beta` coefficient is [11.29, 45.05] as we can see from the table below.

	2.5 %	97.5 %
(Intercept)	-0.974094	0.7311982
nacc_neutral_beta	11.292804	45.0456063

The following plot shows the bootstap distribution of our coefficient, with our bootstrapped confidence interval indicated.



When calculating our bootstrapped 95% confidence interval, we run into an error in many of our bootstrapped samples where "fitted probabilities numerically 0 or 1 occurred". We believe this occurs because we have perfect linear separation in some of our bootstrapped samples, which may occur since our n is quite small. This is interesting because it means that in some of our samples, our coefficient perfectly separates all cases of relapsers from non-relapsers, but this may simply occur as a product of bootstrapping on a small sample of data. Also as a result of that, our model converges to very extreme coefficient values, making it difficult to interpret confidence intervals.

To deal with this problem, we attempted to filter out some of the most extreme values (coefficient values >= 100), and made the plot of the 95% confidence interval (i.e., [1.8228594, 49.3207406]) above. The plot shows a larger 95% confidence interval than our standard glm output, which you'd expect because of the extreme values created by the perfect linear separation issue mentioned above.

(d)

term	estimate	std.error	statistic	p.value
(Intercept)	-49.68342	85519.44	-0.0005810	0.9995365
nacc_drugs_beta	1209.89346	1225903.58	0.0009869	0.9992125
nacc_food_beta	-1950.76984	2071865.39	-0.0009416	0.9992487
nacc_neutral_beta	1976.17890	2023951.49	0.0009764	0.9992209
mpfc_drugs_beta	-247.70075	1101855.62	-0.0002248	0.9998206
mpfc_food_beta	736.55535	1146329.69	0.0006425	0.9994873
mpfc_neutral_beta	-167.65361	650452.85	-0.0002577	0.9997943
vta_drugs_beta	333.21451	1639399.72	0.0002033	0.9998378
vta_food_beta	-443.88852	1069001.80	-0.0004152	0.9996687
vta_neutral_beta	439.08298	1668977.94	0.0002631	0.9997901

term	estimate	std.error	statistic	p.value
acc_drugs_beta	1241.09784	1331858.90	0.0009319	0.9992565
acc_food_beta	100.08731	1833584.41	0.0000546	0.9999564
acc_neutral_beta	-1201.29501	2237031.11	-0.0005370	0.9995715
ains_drugs_beta	-735.34557	2187084.12	-0.0003362	0.9997317
ains_food_beta	-275.56676	3448176.99	-0.0000799	0.9999362
ains_neutral_beta	1123.92079	1668063.10	0.0006738	0.9994624

We again see the perfect linear separation issue mentioned above when working with all the different beta values. This demonstrates why we took care to build parsimonious models in part 2 and 3, and that when working with small data, you must take a lot of care when adding features.

To see whether our significant coefficients changed on a larger model, we also created a model that includes the "neutral betas" for all brain regions in our data (the neutal betas are measure of activity when the subject is presented with neutral cues):

term	estimate	std.error	statistic	p.value
(Intercept)	-0.8693243	0.8017667	-1.0842609	0.2782491
nacc_neutral_beta	34.2791875	12.7161963	2.6957108	0.0070239
mpfc_neutral_beta	-2.0959888	2.7916256	-0.7508130	0.4527652
vta_neutral_beta	4.9032603	3.8896160	1.2606027	0.2074520
acc_neutral_beta	-3.3246783	5.3258159	-0.6242571	0.5324588
ains_neutral_beta	1.2982501	6.6505526	0.1952094	0.8452291

We can see here that `nacc_neutral_beta` is still significant but its p-value is larger (and its coefficient is smaller) than the original single variable glm above. This makes sense when considering the variables we added are activations during the same cues in different brain regions (e.g., `ains_neutral_beta` are the betas extracted from the anterior insula, a brain region thought to be involved in emotion) and are likely to be correlated. Based on prior knowledge of how these regions are connected, you would expect, for instance, the NAcc and Anterior Insula to be correlated. Indeed, the correlation is 0.4.

(e)

Post selection inference may be a problem here. We selected our model after looking at our data. We chose the model that looked like it fit the data best and gave us a low estimated test error. However, when we test for significance, we do not take this selection process into account. This is likely to favorably bias our p-values. One way we could have dealt with this is by creating b bootstrapped samples, using each to sample to create a model, and then looking at how often our coefficient ended up being significant. This would take our model selection process into account. Another way to dealing with post-selection inference would be to validate on new data. Since we didn't have a test set, we couldn't do this.

Since we only have one covariate, multiple hypothesis testing is generally not a problem and no corrections are needed. If our coefficient is actually zero (i.e., the null is true), there is only a 5% chance it will be significant at the $\alpha = .05$ level. However, when we used 10-fold cross-validation to get an estimate of how our model performs on test data, we did conduct multiple (10) hypotheses. We should therefore expect 5% of our rejections to be false positives. If we applied a Bonferroni correction here, our probability of declaring even one false positive would be no more than 5%. We didn't think it was particurly relevant, however, to apply a correction in this case, since our reasons for doing cross-validation in part (b) was just to do something equivalent to fitting our model on a test set.

In part (d), we showed that `nacc_neutral_beta` is correlated `ains_neutral_beta` (in other words, NAcc

activity is correlated with Anterior Insula activity when a person is presented with neutral cues). However, our model only has one coefficient and so this correlation should not affect our p-values.

(f)

It doesn't make sense to interpret the relationship between NAcc activity when looking at neutral cues and relapse in 6 months as causal. Addiction is the underlying mechanism increasing chance of relapse, and there are many complex factors related to addiction that may be confouding our ability to infer causality. For instance, various emotional, personality, and socio-economic factors play a role in addiction, so the variations in these factors may be creating confounds for relapse. We also know (as discussed above) that NAcc activity is correlated with activity in other brain regions. Therefore, we cannot interpret the relationship we discovered causally.

We were not expecting to find a causal relationship, and causal relationship would not make much sense in this context. Our goal was to find a reliable neural signal of risk for relapse within 6 months.

Discussion

These models could be used for both predication and inference. The models could be used by clinicians to predict if a stimulant-dependent veteran will relapse after rehabilitation. The models could also be used to make inferences about the neuroscience of addiction.

Our models should hold up well over time. There's no obvious reason to suspect that time will affect the relationship between nucleus accumbens activity and relapsing. However, updating our data and then refitting the regression model after more time has passed would be useful. This is because our current analysis removes those who hadn't relapsed at the time the data set was completed. However, this doesn't mean that they never relapsed. It would be useful to continually follow-up with them (if possible), update our data, and then refit our regression model. Note that this isn't relevant to our current classification model, since that predicts relapse in 6 months.

Users of our models (e.g., clinicians helping stimulant addicts) should be aware that are data set was small and therefore our models are vulnerable to overfitting. Although we used cross-validation to get an estimate of our test error, this is likely still an underestimate of the true test error. Therefore, users of the model should be cautious about making important decisions (e.g., whether or not to give a patient extra treatment or to release them) using our model, and should definitely not rely solely on our models' predictions.

If we could recollect our data, we would try to increase the number of patients in the study. We would also increase the number of brain regions from which activity was recorded. Our current data includes covariates collected during follow-ups (e.g., obstime, relapse). We would add a survey to assess general well-being to the covariates collected during the follow-up.

If we were to analyse the same data set again, we would try a) using all by-trial data to create a hierarchical model (our current models just use our by-subject data set) and b) use survival analysis to model time to relapse.