Hi Penny,

The reason why you can’t use the hypothesis test and ordinary least square regression is that the data you are using is of high dimensional, say, the p, which is the number of columns, is much larger than the n, which is the number of observations. In this data, we have 10k columns but only 70 observations are provided. Therefore, the information contained in the data does not suffice to explain such a large model provided by ordinary least square regression, and simple hypothesis test is also not suitable here since the instance we care about is whether all the beta are zeros, which involves the simultaneous testing of more than one hypothesis.

So, if we are going to use a hypothesis test only, say, we “bound” the p value of one hypothesis by alpha, which is usually taken as 0.05. Then, an “inflation” is incurred in the overall p value, since rejecting each null hypothesis at level alpha does not guarantee the overall hypothesis to be rejected at level alpha. More intuitively, if we reject each hypothesis at level alpha, actually we are too strict, and we reject many null hypotheses which in fact we should accept. For example, if all the betas are N(0,1) distributed, we still would reject some of the betas to be zero if we use the above method. Therefore, to account for such “inflation”, we need some correction to those p values, and multiple testing does this job for us. By using multiple testing, we reject each hypothesis in a loose sense(usually set the cutoff value to be smaller than alpha so that less hypotheses are rejected in this sense), ensuring that the simultaneous testing as a whole is rejected at level alpha, and we control the false discovery rate, say, we don’t reject too many hypotheses than we should.

Similarly, if we are going to use standard regression in this case, then the design matrix, which you learned from the regression classes, is sparse, or more intuitively, “zero-inflated”, which means that there are many zeros in the entries of the matrix. This indicates that the information we have on hand is not enough to explain all the betas that are proposed in a standard regression model. In other words, many betas tend to be zero in this case, and the data is correlated and can cause unstable solutions to the value of betas. Therefore, we may consider penalized regression like LASSO regression, in which the zero betas are identified and shrunk to zero by adding a penalized term to the loss function of standard regression that you are familiar with. We can view LASSO in the sense of variable selection, and it can provide results which are stable to our data and accounting for the correlation structure in the fMRI scanning.

To sum up, you can use both multiple testing and penalized regression to do the job, but there are some differences between these two methods. Multiple testing usually requires you to propose some p values first, and you can either do this according to some assumptions of the model, for example, use t test to obtain the p values first, or using the non-parametric method. Then, we modify the p values and control the significance level and false discovery rate for the overall hypothesis. Penalized regression, on the other hand, can account for the sparsity of the data and explain the correlation structure by shrinking some betas to zero. However, penalized regression is not suitable for controlling the false discovery rate.

Hope this will help.

Best,

Hongfan