**Ordinary Least Squares, Ridge, and Lasso in Disease Risk Prediction**

Joshua Cheung

**Introduction**

Genome-wide association studies (GWAS) have discovered many single nucleotide polymorphisms (SNPs) associated with complex traits, such as height (Yang et al., 2010). The SNP genotype data from GWAS can be used in risk prediction of these traits (Kooperberg et al., 2010). Predicting risk is oftentimes difficult, since complex traits are usually affected by many different genes, each of which has a small effect on the trait (Wray et al., 2007). Although some question the effectiveness of prediction of complex traits, others see risk prediction as a promising step towards personalized medicine (Janssens et al., 2006).

There have been several methods applied to risk prediction. Several studies employed support vector machines (SVM) in clinical disease risk prediction. In one study, an SVM was applied to a GWAS of Parkinson’s Disease (PD); although the model did not produce strong predictions, the study suggested that inclusion of low-frequency (<5%) variants may be important in risk prediction of PD (Mittag et al., 2012). Another study used an SVM in risk prediction of type 1 diabetes using several hundreds of markers, producing good predictions (Wei et al., 2009). Simple ordinary least squares (OLS) regression can also be used to estimate effect sizes of each SNP on a given trait; more commonly, regularized regression techniques such as lasso and elastic net, which impose penalty functions on the least squares model, have also been used to model risk (Waldmann et al., 2013).

In this project, I aimed to evaluate OLS, along with ridge and lasso regularization in risk prediction. I examined the effect of heritability on the fit of the OLS model, hypothesizing that heritability limits the accuracy of predictions. I also tested the effect of sample size versus the number of SNPs , as other studies have suggested that trait predictions should improve as sample size increases (Wray et al., 2013). To approach this problem, I simulated genotype and phenotype data and used OLS, ridge, and lasso to try to predict phenotypes from genotypes, followed by evaluation of the predictions produced by these methods. The data confirmed both hypotheses, showing that limits strength of predictions and that a high to ratio is required for model accuracy. The data also showed a modified OLS output (described below) and lasso to be good models for risk prediction, while ridge produced a less accurate model.

**Methods**

**Simulations:** In simulations of genotypes and phenotypes, I examined three variables: sample size n, number of SNPs m, and heritability (Zaitlen and Kraft, 2012). For this project, I simulated 3 different sets of data: , , ; , , ; and , , . For each set, I simulated training and validation subsets, each with sample size . Throughout all simulations, 1% of SNPs were causal (had effect size ). I chose a linear model, which produces predictions for continuous traits (such as height).

First, minor allele frequencies were generated uniformly between 0.05 and 0.5. SNPs were simulated for each individual, drawing from the binomial distribution . The causal SNP positions were generated randomly; at these positions, the effect size was simulated from the normal distribution . At non-causal positions, was set to 0. The environmental (non-genetic) contribution was simulated from the normal distribution ), where . Finally, the phenotypes were calculated as the sum of linear effects . values of zero therefore do not contribute to the value of .

**Regression Models:** The first regression model I tested was the OLS model. It estimates , but , where is the Moore-Penrose pseudoinverse of ; this gives .

I also tested modifying the OLS output: if we know that 1% of SNPs are causal for this trait, how would we change our model? In the modified OLS, I sorted the values by decreasing absolute value, setting all but the first 2% to 0. I chose 2% in order to be more inclusive, increasing the likelihood that the model includes the truly causal SNPs.

I also considered two regularization techniques, ridge and lasso. Ridge regression estimates by solving , or in matrix form, , where is some penalty parameter. Ridge regression shrinks values towards each other and is ideal if there are many non-zero predictors (Waldmann et al., 2013). Lasso is similar to ridge, except the penalty term uses the norm of : ). Lasso sets less relevant values to 0, providing an alternative to the modified OLS output (Roth, 2004). The of the lasso was chosen such that 2% of were non-zero; the same was applied for ridge.

The values were calculated using the training set; then, the estimated phenotypes of the validation set were calculated, where .

The models were evaluated for fit in terms of and mean squared error (MSE).

**Results**

**Table 1: Fit of OLS and Regularized Regression Models**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Set 1: , ,** | | | **Set 2: , ,** | | |
|  |  | **MSE** |  |  | **MSE** |
| **OLS** | 0.1817 | 0.6034 | **OLS** | 0.0924 | 0.4566 |
| **Modified OLS** | 0.1742 | 0.6089 | **Modified OLS** | 0.0059 | 0.8906 |
|  | | | | | |
| **Set 3: , ,** | | | | | |
|  |  | **MSE** |  |  | **MSE** |
| **OLS** | 0.8942 | 0.0467 | **Ridge** | 0.6669 | 0.1472 |
| **Modified OLS** | 0.8931 | 0.0472 | **Lasso** | 0.8930 | 0.0472 |

The calculated and MSE values for each model in each data set are given in Table 1. In Set 1 (high , low , low ), the OLS and modified OLS models produced low (0.1817, 0.1742) and high MSE (0.6034, 0.6089). Because there was low fit for this data set, I did not perform regularization on this data set, due to low likelihood of any large improvement of fit. Set 2 (low , high , high ) also produced low (0.0924, 0.0059) and high MSE (0.4566, 0.8906). For the same reason as given for Set 1, regularization methods were not tested on Set 2. In Set 3 (high , low , high ), the OLS mode produced high (0.8942) and low MSE (0.0467). Using ridge () resulted in lower (0.6669) and higher error (0.1472). However, the modified OLS and lasso () both resulted in only slightly lower (0.8931, 0.8930) and slightly higher MSE (0.0472 in both cases). The vs. plots for Set 3 are shown in Figures 1-4. Finally, modifying the OLS output to include only causal was 0.8935, suggesting an upper bound for prediction strength.

**Figure 1: OLS,**



**Figure 3: Ridge,**



**Figure 2: Modified OLS,**



**Figure 4: Lasso,**



**Discussion**

Performance of OLS on Set 1, in comparison to its performance on Set 3 (in which only differed) showed that even with high sample size, a low limits the strength of predictions, as expected – low means that the variation in phenotype is largely due to variation in the environmental contribution, which is not captured by the model. Set 2, in comparison to Set 3 (in which is the reciprocal of that of Set 2) demonstrated that with high , a low to ratio will still lead to weak predictions. Increasing sample size would thus improve model predictions.

Ridge resulted in a reduction of fit; this may be due to the shrinkage of good predictors, but may also be due to the chosen value. The modified OLS and lasso, however, produced almost as good a fit as the OLS model, with the bonus of ignoring 98% of the data. These methods thus have the potential to produce accurate models while reducing number of computations required.

To extend this project, I would try to determine optimal values for ridge and lasso using cross-validation. In addition to these regularization methods, I could investigate the strength of elastic net. Another possibility would be to incorporate environmental effects into the model. Finally, I would test the effect of more intermediate values on model strength. If simulations with intermediate values produce a good fit, the methods described could be applied to real data.

**References**

Janssens, A.C., Aulchenko, Y.S., Elefante, S., Borsboom, G.J., Steyerberg, E.W., and van Duijn, C.M. 2006. Predictive testing for complex diseases using multiple genes: Fact or fiction? *Genet. Med.* 8:395-400.

Kooperberg, C., LeBlanc, M., and Obenchain, V. 2010. Risk Prediction using Genome-Wise Association Studies. *Genet. Epidemiol.* 34: 643-652.

Mittag, F., Büchel, F., Saad, M., Jahn, A., Schulte, C., Bochdanovits, Z., et al., for the International Parkinson’s Disease Genomics Consortium (IPDGC). 2012. Use of Support Vector Machines for Disease Risk Prediction in Genome-Wide Association Studies: Concerns and Opportunities. *Hum. Mutat.* 33: 1708-1718.

Roth, V. 2004. The Generalized LASSO. *IEEE Trans Neural Netw.* 15: 16-28.

Waldmann, P., Mészáros, G., Gredler, B. Fuerst, C., and Sölkner, J. 2013. Evaluation of the lasso and the elastic net in genome-wide association studies. *Front. Genet.* 4: 270.

Wei, Z., Wang, K., Qu, H.Q., Zhang, H., Bradfield, J., Kim, C., et al. 2009. From Disease Association to Risk Assessment: An Optimistic View from Genome-Wide Association Studies on Type 1 Diabetes. *PLoS Genet.* 5:e1000678.

Wray, N.R., Goddard, M.E., and Visscher, P.M. 2007. Prediction of individual genetic risk to disease from genome-wide association studies. *Genome Res.* 17: 1520-1528.

Wray, N.R., Yang, J., Hayes, B.J., Price, A.L., Goddard, M.E., and Visscher, P.M. 2013. Pitfalls of predicting complex traits from SNPs. *Nat. Rev. Genet.* 14: 507-515.

Yang, J., Benyamin, B., McEvoy, B.P., Gordon, S. Henders, A.K., Nyholt, D.R., et al. 2010. Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* 42: 565-569.

Zaitlen, N. and Kraft, P. 2012. Heritability in the genome-wide association era. *Hum. Genet.* 131: 1655-1664.