Appendix

The material contained in this appendix is supplemental information for the manuscript entitled, "Multivariate adaptive regression splines: A powerful method for detecting disease-risk relationship differences among subgroups (York TP, Eaves LJ, van den Oord EJCG)." The first section gives details of the calculations used to specify the parameter values for each simulated condition. The tables in the second section report the precision estimates for the mean squared error, bias, and standard deviation for each simulated condition.

1. Calculations.

To determine the parameter values that would imply the desired proportions of variance explained by the genotype ($r^2_{Y.G}$) and covariate ($r^2_{Y.X}$), we chose values for all parameters of the function $f(X,G)$, except for one genetic effect. The value of this genetic parameter ensures that the proportion of variance explained by the gene and covariate equals the implied ratio $c = r^2_{Y.G} / r^2_{Y.X}$ is then the solution to the equation:

$$c\mathrm{VAR}_{Y.G} = \mathrm{VAR}_{Y.X} \qquad (1)$$

The expressions for $\mathrm{VAR}_{Y.G}$ and $\mathrm{VAR}_{Y.X}$ are given below. We solved this equation numerically using Brent's method (Brent, 1973). Thus, this method basically tries different values for the remaining genetic parameter until the desired ratio c is obtained. There may not always be a solution for the chosen set of fixed parameters so that in practice it may be necessary to try multiple sets. Finally, because the proportion of variance explained by genotype and covariate were specified, the proportion of error variance is also determined and equal to 1- ($r^2_{Y.G} + r^2_{Y.X}$). Using $\mathrm{VAR}(Y)_G$ and $\mathrm{VAR}(Y)_G$ as obtained when solving Equation 1, the total error variance can be derived:

$$\mathrm{VAR}(E) = ((1- r^2) / r^2)(\mathrm{VAR}(Y)_G + \mathrm{VAR}(Y)_G) \quad (2)$$

where $r^2 = r^2_{Y.G} + r^2_{Y.X}$.

We want to express the variances explained by the gene and covariate as a function of the parameters so that Equation 1 can be solved for one of the genetic parameters. Define f_j as the frequency of genotype group j where $j = 0..2$ corresponds with the number of copies of allele A. To compute these genotype frequencies we assume Hardy-Weinberg equilibrium. The variance of dependent variable Y for a given value of the covariate $X = x$ equals:

$$\mathrm{VAR}(Y|X=x) = \sum_{j=0}^{2} f_j(\mathrm{E}(Y|\ X=x,\ G=j) - \mathrm{E}(Y|(X=x))^2 \quad (3)$$

$$\text{with } \mathrm{E}(Y|(X=x) = \sum_{j=0}^{2} f_j \mathrm{E}(Y|\ X=x,\ G=j) \quad (4)$$

The expected values depend on the choice of the function $f(X,G)$ for which, as discussed in the text, we chose different forms. If the covariate has density $f(x)$, the total variance explained by the genotypes becomes:

$$\mathrm{VAR}_{Y.G} = \int_{-\infty}^{\infty} f(x)\mathrm{VAR}(Y|X=x)\ dx \quad (5)$$

Thus, the variance explained by the genotype is basically a weighted sum of the variance at each value of X with weights equal to the "proportion" of the individuals with covariate value $X = x$.

The mean of Y in the whole sample is:

$$\mathrm{E}(Y) = \int_{-\infty}^{\infty} f(x)\mathrm{E}(Y|X=x)dx \quad (6)$$

This overall mean can be used to compute the variance explained by the covariate:

$$\mathrm{VAR}_{\mathrm{Y.X}} = \int_{-\infty}^{\infty} f(x)(\mathrm{E}(Y|X{=}x) - \mathrm{E}(Y))^2 dx \qquad (7)$$

Thus, $\mathrm{VAR}(Y)_X$ is a weighted sum of the squared deviation of the mean for each value of x from the overall mean:

Finally, the proportion of error variance is equal to 1- r^2, where $r^2 = r^2_{\mathrm{y.G1}} + r^2_{\mathrm{y.x}}$. Using $\sigma^2_{\mathrm{y.G1}}$ and $\sigma^2_{\mathrm{y.x}}$ as obtained when solving Equation 1, the total error variance can be derived:

$$\mathrm{VAR}(E) = (r\mathrm{VAR}(Y)_G + \mathrm{VAR}(\mathrm{Y})_X) \times (1 - r^2 / r^2) \qquad (8)$$

The mean squared error for 1,000 replicates of each condition was calculated as:

$$\left[\sum_{j=1}^{1000} \sum_{i=1}^{1200} w_i (obs_i - \exp_i)^2 \Big/ 1000\right] \qquad (9)$$

where j is the number of replicates and i is the number of increments in the covariate range (x-axis). w_i is the normal probability density function used to weight each increment similar to the distribution of the covariate. *obs* is the predicted value from MARS. *exp* is the model value.

Bias for 1,000 replicates of each condition was calculated as:

$$\sum_{j=1}^{1000} \sum_{i=343}^{858} abs(obs_i - \exp_i) \Big/ 1000 \qquad (10)$$

where j is the number of replicates and i is the number of increments in the covariate range (x-axis). *obs* is the predicted value from MARS. *exp* is the model value.

2. Supplemental tables.

Table 1. Estimated fold difference of mean squared error for MARS and polynomials.

Reported as $\sqrt{MARS_{MSE}} \Big/ \sqrt{Polynomial_{MSE}}$.

		Frequency of a allele					
		20%		50%		80%	
		Variance explained by gene					
Model	Sample	2.5%	5%	2.5%	5%	2.5%	5%
Linear	500	4.596	6.481	3.023	2.403	2.265	2.287
	1000	4.312	3.820	5.961	9.547	5.851	7.316
Logistic	500	2.428	2.697	4.052	4.394	5.378	6.829
	1000	3.097	3.633	5.723	5.994	7.561	9.297

Table 2. Estimated fold difference of bias estimation for MARS and polynomials.

Reported as $\sqrt{MARS_{BIAS}} \Big/ \sqrt{Polynomial_{BIAS}}$.

		Frequency of a allele					
		20%		50%		80%	
		Variance explained by gene					
Model	Sample	2.5%	5%	2.5%	5%	2.5%	5%
Linear	500	4.471	6.444	2.062	1.862	1.443	1.426
	1000	3.570	3.246	5.361	8.047	4.271	6.115
Logistic	500	1.862	2.061	2.710	3.306	3.695	5.548
	1000	2.204	2.534	3.544	4.462	5.391	7.742

Table 3. Estimated fold difference of standard deviation estimation for MARS and polynomials. Reported as $\sqrt{MARS_{STDV}} \Big/ \sqrt{Polynomial_{STDV}}$.

		Frequency of a allele					
		20%		50%		80%	
		Variance explained by gene					
Model	Sample	2.5%	5%	2.5%	5%	2.5%	5%
Linear	500	3.535	5.152	4.565	4.326	4.098	5.403
	1000	8.209	9.325	5.291	8.609	5.490	6.778
Logistic	500	2.648	3.044	4.516	4.930	5.649	6.948
	1000	3.605	4.408	7.007	7.365	8.233	9.813

References

Brent RP (1973) Algorithms for minimization without derivatives. Englewood Cliffs, NJ, Prentice-Hall.