Appendix 2

Demonstration of the DSA Model Selection Process

In this demonstration, we apply the DSA procedure, cvGLM( ) –the model selection component of the cvDSA package for conditional means (e.g., treatment model)--using R software. Note that the DSA package can be used as an alternative to cvGLM. This demonstration illustrates the steps by which the DSA algorithm selects the same model that was used to simulate a given data distribution. We simulated data (n=2000 observations, 4 variables x1-x4 from the random uniform distribution), and generated an underlying data distribution based on a simple model

$$\boldsymbol{Y = x1 + x2 + x1x3 + \varepsilon} \qquad (1)$$

where ε ~ Normal(0,0.25). The following parameters were submitted to the cvGLM( ) procedure:

cvGLM(y=Y, x=X, family = 'gaussian', yx.model = list(Model = NULL,
Order = (1, 1, 1, 1), Size = 4, Interaction = 2, ncv=2)

Y equals the vector of responses (y1,…,yn) based on the simple model (1) given above; X represents the design matrix of the combined uniform random variables x1-x3, as well as x4 which was not used in the data-generating model for Y; family='gaussian' indicates that the outcome variable Y is continuous and the error has a Gaussian distribution; ncv = 2 indicates 2-fold cross-validation; and 'yx.model' is a list object which consists of 4 elements:

1) Model = NULL indicates that the user expects the function to return the appropriate model that relates **y** and **X.** Alternatively, the user can supply a model, and the function will fit the coefficients.
2) Order = c(1,1,1,1) is a numerical vector that indicates that the power of each of

the four variables in the model should not exceed 1;

3) Size = 4 specifies the maximal number of terms in the model;
4) Interaction = 2 indicates the model only allows main terms and 2-way interactions.

The simulation model used to generate the data was

$$\mathbf{-1 + (1)x_1+(1)x_2+(1)x_1x_3}$$

and the algorithm returned the estimated model

$$\mathbf{-1.014 + (0.991)x_1+(1.023)x_2+(1.002)*x_1x_3}$$

The cross-validation risk associated with this latter model (0.063) (see below) measures the discrepancy between the fit of the best size 3 model to the simulated data. Given the relative comparability of the estimates of the parameters to the true parameter values, it is clear that most of difference can be attributed to the random noise incorporated as part of the simulation and not the result of systematic error of DSA to identify the correct model. The steps used by the algorithm to generate models and select an optimal model of the data distribution (i.e., model used to simulate the data) have been extracted from output and provided below.

Depiction of the DSA model selection process as applied to simulated data in Appendix 2.

```
Intercept
Add     { x1, x2, x3, x4 }
        Select   x1
        Substitution    { x1x2, x1x3, x1x4 }
                Select          x1x2
                Substitution    { x2, x2x3, x1x3, x2x4, x1x4, x1 }
                Add             x1x2 + { x1, x2, x3, x4, x2x3, x1x3, x2x4, x1x4 }
                        Select          x1x2 + x1x3
                        Delete          { x1x2, x1x3 }
                        Substitute      x1x2 + { x3, x1x4, x2x3, x3x4, x1}
                                        x1x3 + { x2, x1x4, x2x3, x2x4, x1 }
                        Add             x1x2 + x1x3 + {x1, x2, x3, x4, x1x4, x2x3, x3x4, x2x4 }
                        Select          x1x2 + x1x3 + x2
                                        Deletion        x1x2 + x1x3, x1x2 + x2, x1x3 + x2
                                        Substitute      x1x2 + x1x3 +  { x2x3, x2x4 }
                                                        x1x2 + x2 + { x1, x3, x2x3, x3x4, x1x4 }
                                                        x1x3 + x2 + { x1, x2x3, x2x4, x1x4 }
                                        Select          x1 + x1x3 + x2
                                                        Delete          x1x3 + x2, x1 + x2, x1 + x1x3
                                                        Substitute      x1 + x1x3 + {x1x2, x2x3, x2x4 }
                                                                        x1 + x2 + { x3, x2x3, x1x2, x3x4, x1x4 }
                                                                        x1x3 + x2 + { x1x2, x1x4 }
                                                        Add             x1 + x1x3 + x2 + { x3, x4, x1x2, x2x3, x2x4, x3x4, x1x4 }
                                                                        Select  x1 + x1x3 + x2 + x3x4
```

Selected model of size 3: x1 + x1x3 + x2.
Final Selected Model: -1.014 + 0.991*x1+1.023*x2+1.001*x1x3   (Cross-validation Risk 0.0629)