Optimal Data Analysis
Vol. 5 (November 4, 2016), 171-174

2155-0182/10/$3.00

Theoretical Aspects of the D Statistic

Paul R. Yarnold, Ph.D., and Ariel Linden, Dr.P.H.
Optimal Data Analysis, LLC Linden Consulting Group, LLC

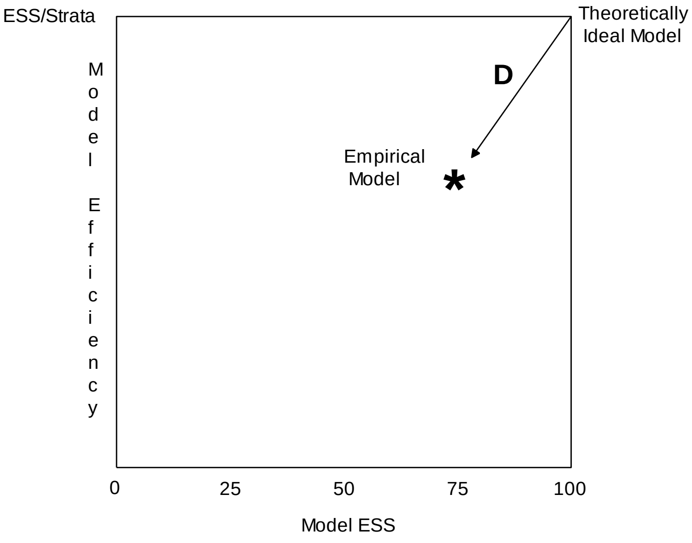
Conceptualization of the D statistic in novometric theory[1] is advanced.

In novometric theory the minimum denominator selection algorithm identifies the descendant family (DF), consisting of one or more optimal models, within which the globally optimal (GO) model for an application resides. In applications with a DF consisting of multiple models, the set of optimal models differ with respect to their *predictive accuracy* normed *vs*. chance by the ESS statistic (0=predictive accuracy expected by chance; 100=errorless prediction), and their *complexity* defined as the number of sample strata (endpoints) in the model. Observations are *homogeneous* with respect to the attributes that define the endpoint *within* a given endpoint, and are *heterogeneous* with respect to the defining attributes *between* model endpoints.[1,2]

Novometric analysis identifies the GO model in the DF representing the “best” combination of predictive accuracy and *parsimony* (the antithesis of complexity), defined as ESS divided by the number of model strata (i.e., the mean ESS-per-endpoint obtained by the model). This assessment requires creating a square Cartesian space, crossing ESS by efficiency as abscissa and ordinate, respectively. In Figure 1, for example, for a two-strata model the efficiency axis ranges from 0 to 100/2=50; for a three-strata model the efficiency axis ranges from 0 to 100/3=33.3; and for a model with *s* strata the efficiency axis ranges from 0 to 100/*s*.

The *distance* of an empirical model from a theoretically ideal model[1] is computed as: D= 100/(ESS/*s*)-*s*. For example, for a model with s=3 and ESS=60, D=100/(60/3)-3=2: two more effects of equivalent efficiency are needed to achieve an ideal statistical solution. As seen, the distance D of an empirical model (asterisk) from a theoretically ideal model (upper right-hand corner) for an application is a function of both accuracy and parsimony of the empirical model. Normed over accuracy and parsimony, D may thus be used to directly compare the quality of competing models regardless of their underlying architecture—each model considered relative to its corresponding theoretically ideal model.

Figure 1: ESS by Complexity Space, the Theoretically Ideal Statistical Model, and D



Copyright 2016 by Optimal Data Analysis, LLC
2155-0182/10/$3.00

Theoretically Ideal Model: An Attainable Upper Bound of D

The accuracy-by-parsimony space in Figure 1 is theoretically bounded by the intersection of the corresponding maximum values (100 and 100/*s*, respectively) of these indexes. Several such ideal models have been identified in applied studies involving small samples, and several studies yielded models with D statistics less than one. Unfortunately the small samples for which ESS=100 lacked sufficient statistical power (violating the first axiom of novometric theory) to determine if continued application of the minimum denominator selection algorithm[1] would identify even more parsimonious perfect models. However, in such a circumstance the most parsimonious perfect model that emerges is selected as the GO model for the application. Conceptually related investigator-determined *a priori* selection heuristics are available in ODA software for selecting among multiple optimal statistical models, if they occur.[1-3] Additional study in this area is warranted.

Theoretically Least-Ideal Model: A Non-Attainable Lower Bound of D

As model ESS and efficiency approach zero, D approaches infinity in the limit since the denominator term, ESS/*s*, is zero when ESS=0 (the first axiom of novometric theory requiring sufficient statistical power inhibits overfitting—manifest in terms of an untenably large *s*). In practice this limit is ordinarily not an issue: in typical samples small values of ESS are not statistically reliable, and D statistics are only computed for statistically reliable models that are identified in the DF. That is, statistical unreliability implies that the existence of the model is unproven. Numerous empirical studies have reported negative ESS values for models evaluated using “leave-one-out” one-sample jackknife cross-generalizability analysis (the fifth axiom of novometric theory mandates successful replication), indicating predictive accuracy worse than expected by chance. However, D statistics are not computed because such models are not statistically reliable.

However, in some applications the desired effect size is ESS=0. For example, in causal inference research “propensity scores” are used in an effort to identify a matching or weighting scheme that equates two or more groups with respect to a set of covariates.[4-8] In this perspective, identification of one or more statistically reliable models indicates “bias” compared to a finding of no statistically reliable model. The level of bias associated with each model in the DF can be assessed using the D statistic. The model with the largest D is least similar to a theoretically ideal model, and is thus used to construct propensity scores for matching or weighting purposes. Comparison of test statistic magnitude to select a model is conceptually similar to Akaike’s and Schwarz’s Bayesian information criteria (AIC and BIC) statistics for which one compares relative, rather than absolute values among a set of models for model-selection purposes.[9,10] However, if instead no statistically reliable model is identified, this is considered to be statistical evidence that propensity-score-based matching or weighting returned groups that, as desired, could not be discriminated with respect to the observed covariates. Additional study in this area is warranted.

D-Statistic Penalty Adjustment for Increasing Model Complexity

Table 2 lists, and Figure 2 illustrates the value of the D statistic for models having two, four, six or eight strata, for ESS values ranging between perfect (100) and relatively weak (10).

Regardless of their underlying complexity, all models converge to D=0 in the limit as ESS=100. As stated earlier, application of the minimum denominator search algorithm to a perfect model identified in the DF may reveal additional perfect models with fewer endpoints, and/or additional perfect models with the same number of endpoints but a greater minimum

Copyright 2016 by Optimal Data Analysis, LLC
2155-0182/10/$3.00

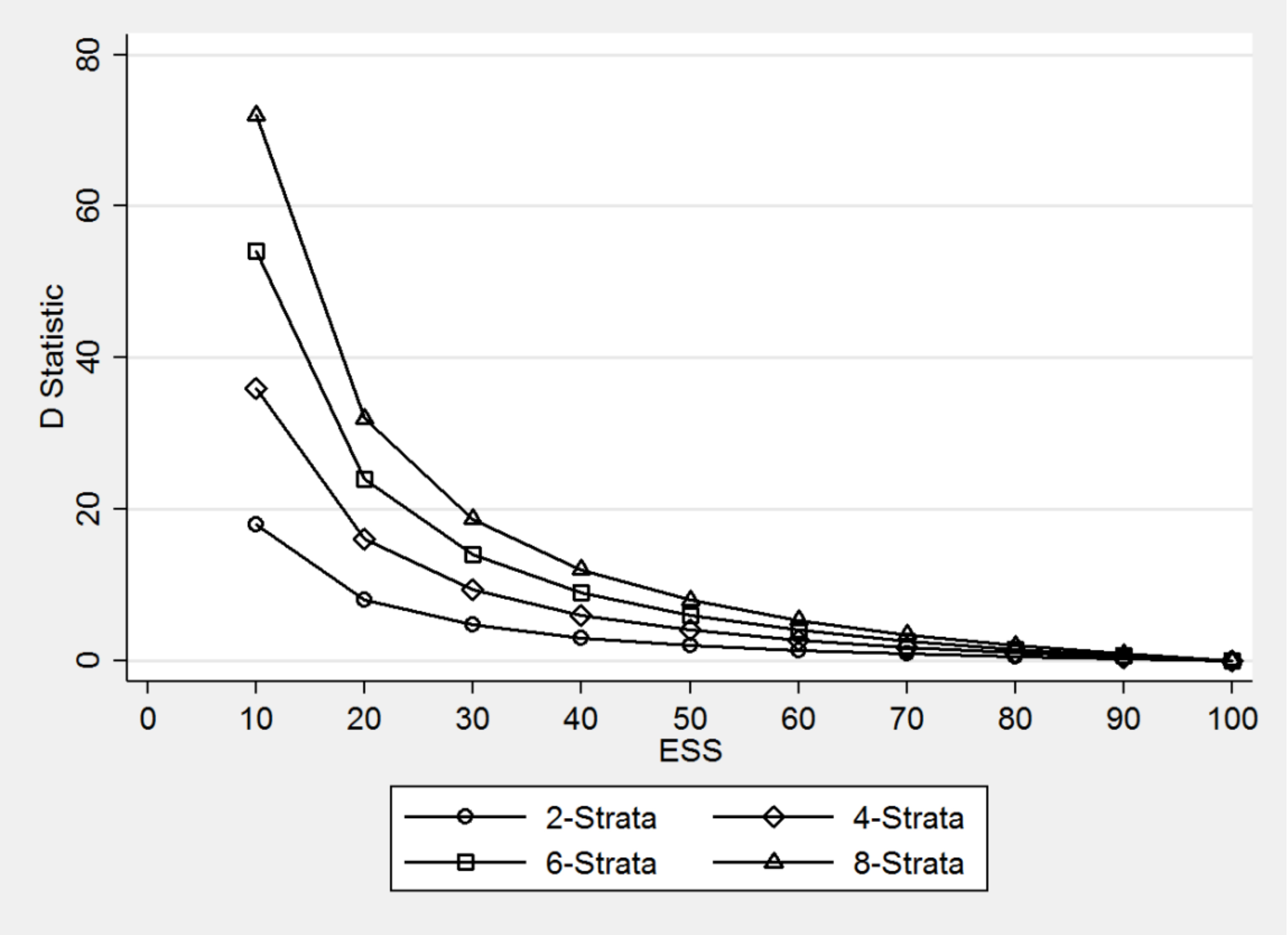
endpoint denominator—thus providing greater statistical power for subsequent analyses.[1]

As clearly seen in Figure 2 the D statistic assigns an increasingly stiffer penalty for model complexity as predictive accuracy diminishes, particularly for effects classified by rule-of-thumb[1,3] as reflecting an effect of moderate or weaker strength (i.e., ESS<50). However, even for very strong models having ESS=90, Table 2 reveals a 400% difference in the corresponding D statistics of the 2- and 8-strata models. To compare the training and validity performance of competing models, exact discrete 95% confidence intervals (CIs) are computed for D vis-à-vis bootstrap analysis, and then the CIs for D are examined for overlap between groups.[1]

Table 2: ESS, Strata, and D

	D Statistic			
ESS	2-Strata	4-Strata	6-Strata	8-Strata
100	0.00	0.00	0.00	0.00
90	0.22	0.44	0.67	0.88
80	0.50	1.00	1.50	2.00
70	0.86	1.71	2.57	3.43
60	1.33	2.67	4.00	5.33
50	2.00	4.00	6.00	8.00
40	3.00	6.00	9.00	12.00
30	4.67	9.33	14.00	18.67
20	8.00	16.00	24.00	32.00
10	18.00	36.00	54.00	72.00

Figure 2: Model Accuracy (ESS), Complexity (Strata), and D



References

[1] Yarnold PR, Soltysik RC (2016). *Maximizing predictive accuracy*. Chicago, IL: ODA Books. DOI: 10.13140/RG.2.1.1368.3286

[2] Yarnold PR (1996). Discriminating geriatric and non-geriatric patients using functional status information: An example of classification tree analysis via UniODA. *Educational and Psychological Measurement*, *56*, 656-667.

[3] Yarnold PR, Soltysik RC (2005) *Optimal data analysis: A Guidebook with Software for Windows*. Washington, DC: APA Books.

[4] Rosenbaum PR, Rubin DB (1983). The central role of propensity score in observational studies for causal effects. *Biometrika*, *70*, 41–55.

[5] Linden A, Yarnold PR (In Press). Combining machine learning and propensity score weighting to estimate causal effects in multi-valued treatments. *Journal of Evaluation in Clinical Practice*.

[6] Linden A, Yarnold PR (In Press). Combining machine learning and matching techniques to improve causal inference in program evaluation. *Journal of Evaluation in Clinical Practice*.

[7] Linden A, Yarnold PR (In Press). Using machine learning to assess covariate balance in matching studies. *Journal of Evaluation in Clinical Practice*. DOI: 10.1111/jep.12538

[8] Linden A, Yarnold PR (In Press). Using data mining techniques to characterize participation in observational studies. *Journal of Evaluation in Clinical Practice*. DOI: 10.1111/jep.12515

[9] Akaike H (1973). Information theory and an extension of the maximum likelihood principle. In: *Second International Symposium on Information Theory*, BN Petrov and F Csaki (Eds.), Budapest: Akailseoniai–Kiudo (pp. 267–281).

[10] Schwarz G (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*, 461–464.

Author Correspondence

Mail: Optimal Data Analysis, LLC
6348 N. Milwaukee Ave., #163
Chicago, IL 60646