Sawtooth Software

RESEARCH PAPER SERIES

Fast, Accurate Driver Analysis

Keith Chrzan
Sawtooth Software

3210 N. Canyon Rd., Provo, Utah
+1 801 477 4700
www.sawtoothsoftware.com

Fast, Accurate Driver Analysis

Keith Chrzan, Sawtooth Software

May 2026

When we run driver analysis, we have to worry that the halo effect is causing our drivers to be correlated with one another (a condition called multicollinearity). To work around this multicollinearity, we do what people call "Shapley regression." Essentially we run our regression analysis, entering each variable in turn and keeping track of the incremental R-squared attributable to the addition of each variable. We do this for every possible ordering, which gets to be a pretty big number of regressions. Sometimes we can simplify things a little by running regressions over millions of combinations instead of trillions of permutations, but still, it gets to be a pretty nasty business.

At the 2026 Sawtooth Research Conference, Dave Lyon showed how to do driver analysis for large problems using "component order of addition" (COA) designs. For a model with k predictors, averaging over all orderings would involve k! regressions, while the COA design reduces that to just k(k-1) regressions – for 20 variables that's the difference between 2.4 quintillion regressions and 380 regressions, respectively. In a previous post I illustrated how much time this saves and how it can open up opportunities to do AOO for different kinds of models (like logit models).

Dave also showed that in his data sets the COA designs produced results that were nearly identical to those you get from running all possible permutations when what you're averaging over orderings is the incremental R-squared attributable to each variable. I wondered how the results of running driver analysis with the COA designs would compare to two other ways of quickly running driver analysis:

- A method called Johnson's relative weights uses a slick bit of linear algebra to run driver analysis quickly without the computational effort of running regression in every possible ordering of variables
- Borrowing from some time I spent doing sequential monadic concept testing I remembered Williams designs – if these worked they would further reduce the number of regressions needed to just 2k (or 40 in the case of 20 variables).

I ran the COA driver analysis on five data sets I had handy and compared it to Johnson's relative weights and to driver analysis using Williams designs. While all three methods produce importances that are extremely similar to what you get when you run all possible orders, sure enough, COA designs do the best, in all five data sets. Johnson's weights and

regression with Williams designs are close to Shapley value regression results but close as they are, the COA designs get closer still. Here's an example from an airline customer satisfaction study (RMSE is "root mean square error" and lower numbers mean that the method is more similar to the Shapley regression result than are higher numbers):

Variable	Shapley	Johnson	COA	Williams
x8	23.59	23.22	23.69	24.51
x2	16.48	16.87	16.60	17.25
x3	10.75	11.04	10.67	11.14
x9	10.20	10.58	10.25	9.79
x5	8.59	8.27	8.45	8.13
x6	8.27	7.81	8.28	7.59
x1	7.82	8.03	7.90	7.16
x7	7.27	7.02	7.24	7.20
x4	7.04	7.16	6.92	7.23
x8	23.59	23.22	23.69	24.51
RMSE vs Shapley		0.542	0.190	0.653

While averaging R-squared over orderings is the most common way to do regression-based driver analysis, there are a couple of other methods, too.

Kruskal's relative importance measure averages incremental squared partial correlation instead of incremental R-squared. Again, however a version of Kruskal's that uses the COA design come closer to averaging over all orderings than does a version using the Williams design:

Variable	Permutations	COA	Williams
x8	25.03	25.12	25.87
x2	17.08	17.17	17.87
x3	10.74	10.67	11.19
x9	10.21	10.27	9.88
x5	8.19	8.03	7.64
x6	7.79	7.83	7.05
x1	7.48	7.57	6.84
x7	6.79	6.79	6.80
x4	6.68	6.56	6.86
RMSE vs All Permutations		0.091	0.571

Note, we don't report Johnson's relative weights for Kruskal's method, because what Johnson's method apportions is R-squared.

We find the same result when we look at Theil's relative importance measure, which measures the drop in an information theoretic entity called "entropy" over orderings: COA gets much closer to averaging over all orderings than does the simpler (and smaller) Williams design:

Variable	All Permutations	COA	Williams
x8	25.42	25.51	26.30
x2	17.11	17.20	17.89
x3	10.65	10.57	11.08
x9	10.09	10.14	9.75
x5	8.19	8.03	7.64
x6	7.80	7.84	7.07
x1	7.40	7.48	6.76
x7	6.76	6.76	6.75
x4	6.60	6.47	6.76
RMSE vs All Permutations		0.092	0.573

I was able to confirm the great performance Dave reported for COA designs, compared to other ways of speeding up driver analysis (Johnson's relative weights and Williams designs). I was also able to confirm that this performance extends to cases where we average things other than incremental R-squareds in our regression analyses – specifically, the incremental squared partial correlations used by Kruskal's importances and the incremental drop in entropy used by Theil's importances.

I use the term "Shapley regression" in the case of the first method, but the name really covers Kruskal's and Theil's methods as well – all three use Shapley's averaging-over-orderings strategy, they just apply it to different aspects of the regression model. If one wanted to be very clear, one might call the first method, the one using incremental R-squared, "LMG," like R does. LMG stands for Lindeman, Merenda and Gold, the authors who first wrote about averaging R-squared over orderings.

References

Kruskal, W. (1987) “Relative importance by averaging over orderings,” *The American Statistician*, **41**, 6-10.

Lindeman, R.H., P.F. Merenda and R.Z. Gold (1980) *Introduction to Bivariate and Multivariate Analysis*. Glenview, IL: Scott, Foresman.

Lyon, D.W. (2026) “Shapley Values for Large Problems,” paper presented at the 2026 Sawtooth Research Conference, Berlin.

Theil, H. and C.F. Chung (1988) “Information-theoretic measures of fit for univariate and multivariate linear regression,” *The American Statistician*, **42**: 249-252.

Williams, E. J. (1949): "Experimental Designs Balanced for the Estimation of Residual Effects of Treatments." *Australian Journal of Scientific Research*, Ser. A, 2, 149–168.