Oct 7th, 1:45 PM - 3:00 PM

All Possible Regressions Using IBM SPSS: A Practitioner's Guide to Automatic Linear Modeling

T. Chris Oshima
Georgia State University, oshima@gsu.edu

Theresa Dell-Ross
Georgia State University, tdellross1@gsu.edu

ALL POSSIBLE REGRESSIONS USING IBM SPSS: A PRACTITIONER'S GUIDE TO AUTOMATIC LINEAR MODELING

T. Chris Oshima
Theresa L. Dell-Ross
Georgia State University

INTRODUCTION

- All possible subsets regression procedure (or all possible regressions) as a preferred method for selecting the "best" model in multiple regression
- May not have been the most frequently used method by SPSS users partly due to its time consuming nature
- Automatic Linear Modeling introduced in Version 19 of IBM SPSS, enabling researchers to select the best subset automatically
- A potential threat of misuse due to its simplicity

PURPOSE

The purpose of this paper is to provide brief information on all possible regressions and to provide a practical guide on how to make the best use of Automatic Linear Modeling.

REGRESSION MODELING

Model selection procedures include:

- Forward selection
- Backward selection
- Stepwise selection

Susceptible to misuse due to software automation

Each procedure is based on its own algorithm in terms of inclusion and exclusion of variables.

REGRESSION MODELING (CONT.)

- Unlike forward selection and backward selection, stepwise regression permits a variable that has been excluded to re-enter into the model.
- Stepwise regression has been subject to criticism. For example, Huberty (1989) advised against the use of the stepwise selection procedures because it is sensitive to the order of variable entry in the model.
- The stepwise regression procedure can overlook other viable subsets.

REGRESSION MODELING (CONT.)

- Huberty (1989) recommended the examination of all possible subsets of predictor variables.
- The only instance in which Huberty recommended the use of the stepwise regression procedure was when the number of predictors was large.
- This procedure challenged investigators in terms of its rigor until the arrival of Automatic Linear Modeling in SPSS Version 19 in 2010.

ALL POSSIBLE REGRESSIONS

- All Possible Regressions as a method that has been strongly endorsed by methodologists
- Most recent textbooks include a section or chapter on all possible regressions (e.g., Chatterjee & Hadi, 2012; Mendenhall & Sincich, 2012; Montgomery, Peck, & Vining, 2012).
- All possible combinations of *k* predictor models are evaluated using various criteria.
- The number of models (2^k models, including the intercept only model) increases rapidly as the number of predictors increases.

ALL POSSIBLE REGRESSIONS (CONT.)

Recommended steps for all possible regressions:

1. Identify all 2^k of the possible regression models and run these regressions.
2. Calculate various criteria for model fit for each model.
3. Evaluate the criteria and come up with model(s) that will answer the research question. This step may include refining the regression equation by transformation and/or adding the interaction terms.

The above process is a daunting task especially if *k* is large.

ALL POSSIBLE REGRESSIONS (CONT.)

- SPSS lacked an automated system until the arrival of Automatic Linear Modeling.
- SPSS syntax could be employed to run all 2^k regression models, but fit indices were not automatically compared.
- However, Automatic Linear Modeling is not a complete process for Steps 1-3. It accomplishes Step 1 and part of Step 2.
- However, the user still had to aggregate various fit criteria manually to make the comparisons possible.

MODEL FIT INDICES

Available indices for evaluating the "best" model include:

- R^2 (The larger, the better)
- Adjusted R^2 (The larger, the better.)
- Mean Square Residual, MSE (The smaller, the better.)
- Adequate R^2 (Choose a model in which R^2 is larger than adequate R^2.)
- Mallow's Cp (Choose a model in which Cp is closest to $k + 1$.)
- Akaike's Criterion Information, AIC (The smaller, the better.)
- Akaike's Criterion Information Corrected, AIC_c (The smaller, the better.)

These indices address different aspects of model fit; therefore it is possible that the choice for "best" model can differ.

MODEL FIT INDICES (CONT.)

Aitkin's Adequate R^2 (Aitkin, 1974)

$$AdequateR^2 = 1-(1-R^2_{FM})(1+(k\cdot F^{*}_{k,N-k-1}/(N-k-1)))$$

Mallow's Cp (Mallow, 1976)

$$C_p = \frac{SSE_{(k)}}{MSE_{(FM)}} - N + 2(k+1)$$

Akaike Criterion Information (AIC) (Akaike, 1974)

$$AIC = N * \ln(\frac{SSE_{(k)}}{N}) + 2(k+1)$$

Akaike Criterion Information Corrected (AIC_c) (Hurvich & Tsai, 1989)

$$AIC\_c = AIC + \frac{2(k+1)(k+2)}{(N-k)}$$

AUTOMATIC LINEAR MODELING IN SPSS

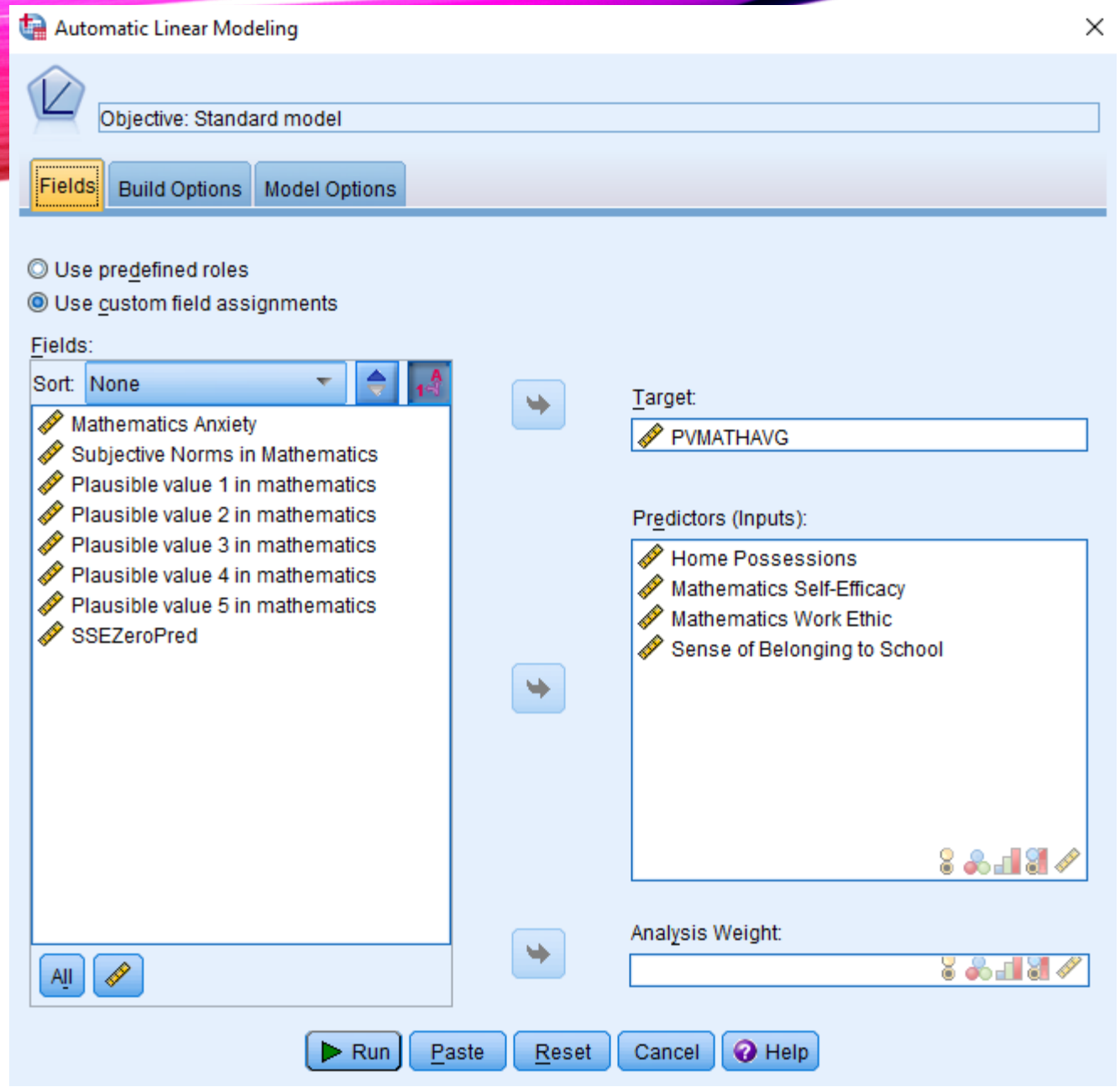
- Automatic Linear Modeling: Analyze > Regression
- Automatic Linear Modeling includes automatic data preparation (ADP) steps.
- One of the model building options is to perform model selection by "Best subsets."
- If $k \leq 20$, then SPSS searches all subsets. If k is larger than 20, a hybrid method that combines the forward stepwise method and the all possible subsets method is performed.
- Three model fit options are available: AIC_c, Adjusted R^2, and Overfit Prevention Criterion (ASE).

DEMONSTRATION

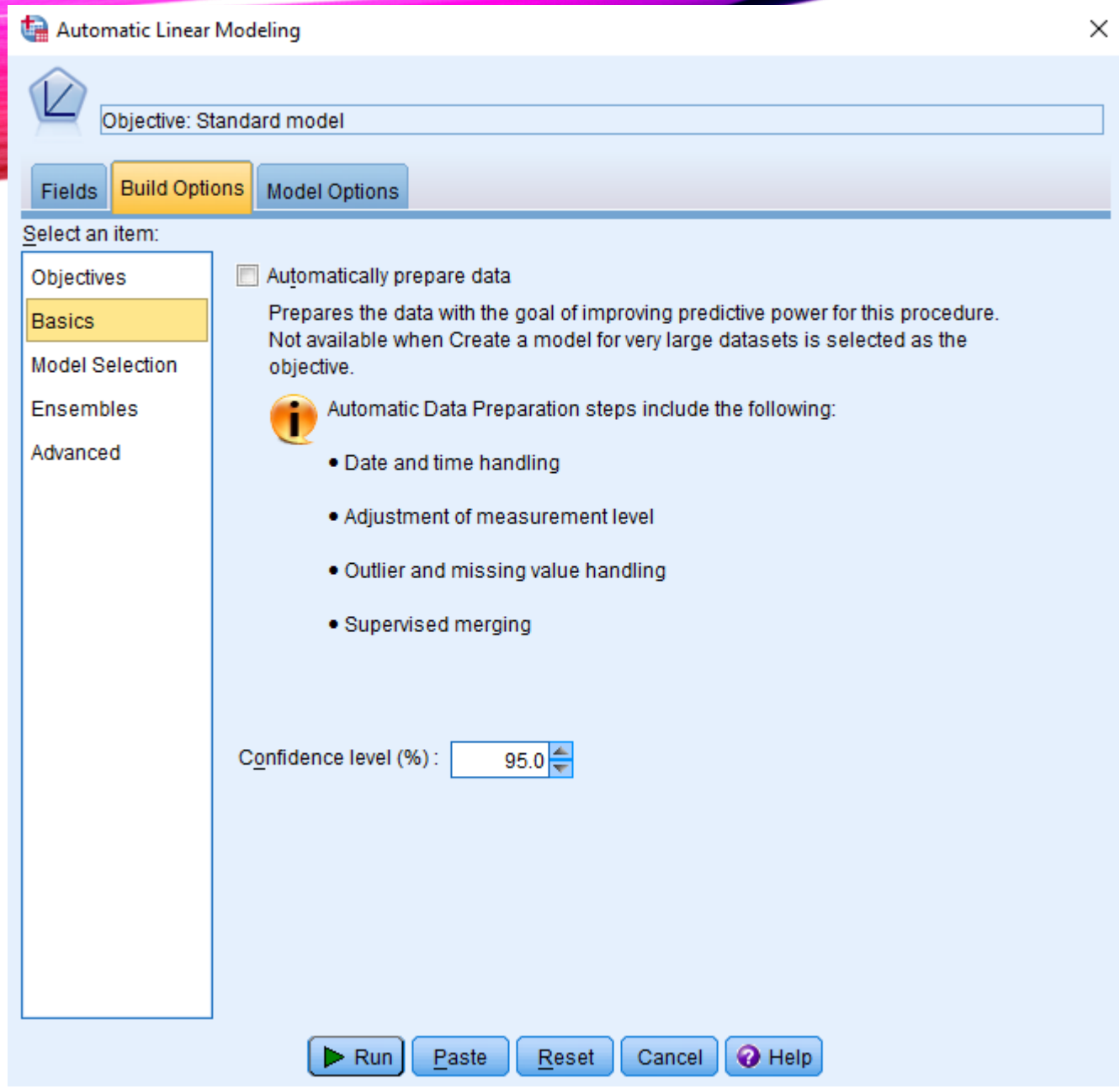
In order to demonstrate how Automatic Linear Modeling can be used effectively, data from the 2012 Program for International Student Assessment (PISA) was analyzed.

- Home Possessions (HOMEPOS)
- Mathematics Self-Efficacy (MATHEFF)
- Mathematics Work Ethic (MATWKETH)
- Sense of Belonging in School (BELONG)

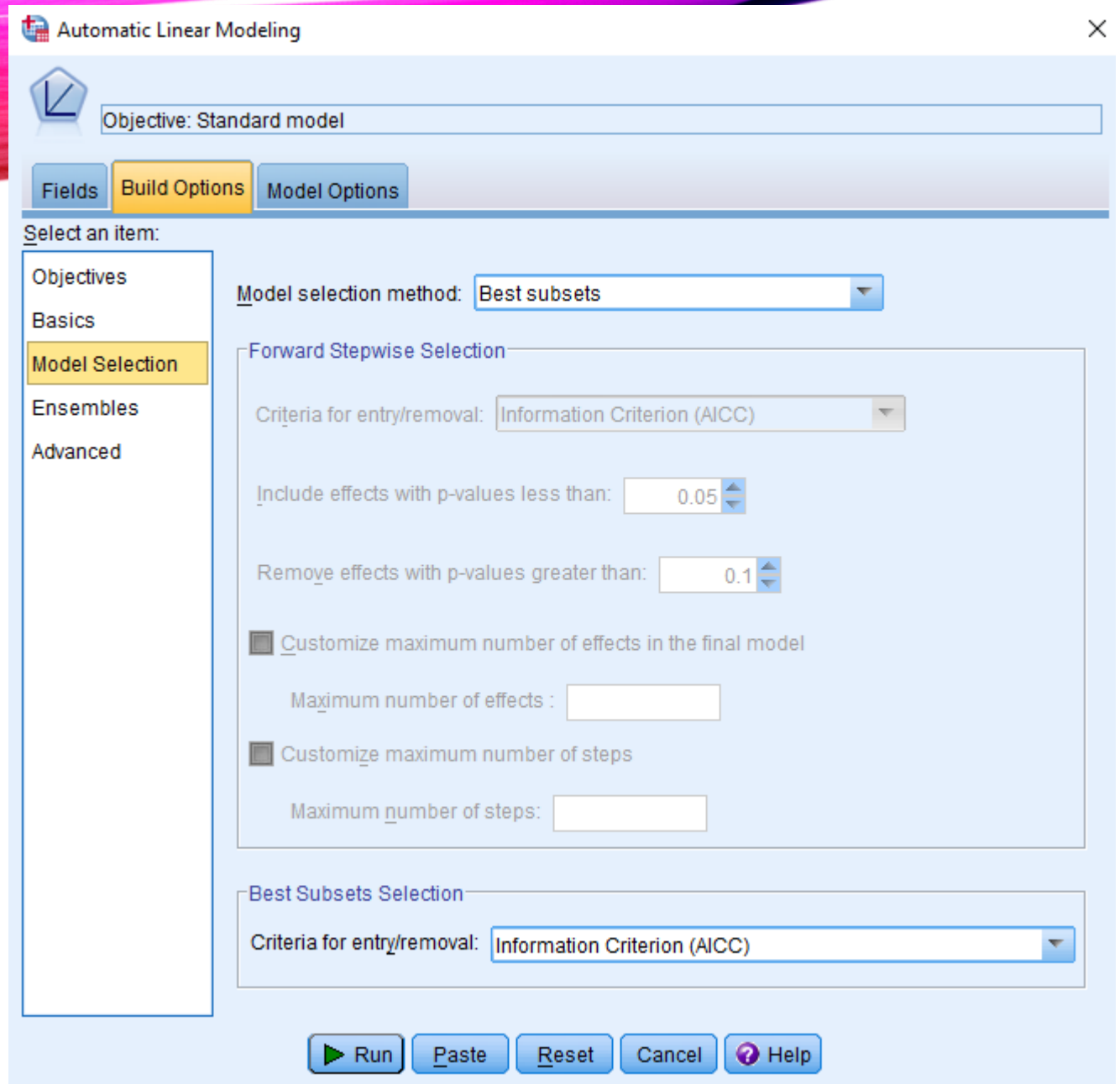
AUTOMATIC LINEAR MODELING IN SPSS (CONT.)



AUTOMATIC LINEAR MODELING IN SPSS (CONT.)



AUTOMATIC LINEAR MODELING IN SPSS (CONT.)



EVALUATION OF AUTOMATIC LINEAR MODELING IN SPSS

		Model									
		1	2	3	4	5	6	7	8	9	10
Information Criterion		13,462.778	13,473.342	13,479.323	13,500.546	13,545.517	13,551.437	13,556.841	13,570.244	13,974.474	13,974.741
Effect	HOMEPOS	✓	✓	✓	✓					✓	✓
	MATHEFF	✓	✓	✓	✓	✓	✓	✓	✓		
	MATWKETH	✓	✓			✓	✓			✓	
	BELONG	✓		✓		✓		✓			

The model building method is Best Subsets using the Information Criterion.
A checkmark means the effect is in the model.

EVALUATION OF AUTOMATIC LINEAR MODELING IN SPSS (CONT.)

# Predictors	Regressors	SSE(k)**	R^2	Adj R^2	MSE(k)	Mallow's Cp***			Akaike's Information		If R^2> Adeq R^2
						Cp	k + 1	|Cp - (k + 1)|	AIC	AIC_c	
0	0	37397291.860	0.000	0.000	23549.932	6259.243	1.000	6258.243	15997.269	15997.272	0
1	X1	32913711.558	0.110	0.110	20739.579	5320.553	2.000	5318.553	15796.340	15796.347	0
1	X2	16406919.944	0.319	0.319	10338.324	1857.299	2.000	1855.299	14690.100	14690.107	0
1	X3	23430298.945	0.010	0.009	14763.893	3330.859	2.000	3328.859	15256.303	15256.311	0
1	X4	24147932.050	0.005	0.004	15216.088	3481.424	2.000	3479.424	15304.241	15304.249	0
2	X1X2	15575532.420	0.354	0.353	9820.638	1684.868	3.000	1681.868	14609.469	14609.484	1
2	X1X3	21030189.459	0.111	0.110	13259.892	2829.297	3.000	2826.297	15086.578	15086.593	0
2	X1X4	21758863.734	0.103	0.102	13719.334	2982.179	3.000	2979.179	15140.703	15140.718	0
2	X2X3	15830593.714	0.329	0.329	9981.459	1738.382	3.000	1735.382	14635.279	14635.294	0
2	X2X4	8307718.056	0.308	0.307	5238.158	160.024	3.000	157.024	13610.756	13610.772	0
2	X3X4	11510699.822	0.009	0.007	7257.692	832.034	3.000	829.034	14128.917	14128.932	0
3	X1X2X3	14933662.458	0.367	0.366	9421.869	1552.199	4.000	1548.199	14544.598	14544.624	1
3	X1X2X4	7930611.979	0.339	0.338	5003.541	82.904	4.000	78.904	13538.940	13538.965	0
3	X1X3X4	10450072.903	0.100	0.098	6593.106	611.507	4.000	607.507	13977.311	13977.336	0
3	X2X3X4	7963378.831	0.314	0.312	5024.214	89.779	4.000	85.779	13545.492	13545.517	0
4	X1X2X3X4	7549767.841	0.349	0.348	4766.268	5.000	5.000	0.000	13462.740	13462.778	1

X1 = HOMEPOS X2 = MATHEFF X3 = MATWKETH X4 = BELONG

DISCUSSION

- Automatic Linear Modeling makes the tedious task of comparing all possible regression models virtually effortless.
- However, the ease of the procedure also presents a danger of having the computer dictate one's research conclusions.
- Automatic Linear Modeling can be an indispensable screening tool especially when there are many predictors. However, once a handful of final candidates are chosen, it is the researcher's responsibility to carefully evaluate those models with various criteria along with substantive questions.

DISCUSSION (CONT.)

Note:

The PISA analysis presented herein was for demonstration purposes only.

These “results” are not meant to contribute to the existing literature on mathematics achievement.

CONTACT

If you are interested in a draft of this working paper, please contact Theresa Dell-Ross (tdellross1@gsu.edu).

REFERENCES

- Aitkin, M. A. (1974). Simultaneous influence and choice of variable subsets in multiple regression. *Technometrics. 16(2),* 221-227.
- Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19 (6*), 716–723.
- Burnham, K. P., & Anderson, D. R. (2002). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach* (2nd ed.). Springer-Verlag.
- Chatterjee, S., & Hadi, A. S. (2012). *Regression analysis by example (5th ed).* John Wiley and Sons, Inc.
- Draper, N. R., & Smith, H. (1981). *Applied regression analysis (2nd ed)*. New York: Wiely.
- Huberty, C. (1989). Problems with stepwise methods-Better alternatives. *Advances in Social Science Methodology. 1*, 43-70.
- Hurvich, C. M., & Tsai, C. L. (1989). Regression and time series model selection in small samples, *Biometrika, 76*, 297–307 .
- Mallow, C. L. (1973). Some comments on Cp. Technometics, 15(4), 661-675.
- Mendenhall, W., & Sincich, T. (2012). *A Second Course in Statistics: Regression Analysis* (7th ed.). Pearson.
- Montgomery, D. C., Peck, E. A., & Vining, C. G. (2012). *Introduction to linear regression analysis*. John Wiley and Sons, Inc.