Model Building Procedures

Researcher Determines Model

Simultaneous and hierarchical model testing approaches are clearly the most common procedures used in psychology and the social sciences. In both methods, the predictors (covariates) are chosen by the researcher based on theory and prior research rather than chosen by some empirical means, such as after examining the strength of their correlations with the outcome. The distinction parallels the distinction we have used in other contexts—theoretically driven *a priori* and empirically driven *post hoc*.

*Simultaneous***.** All predictor variables are entered at the same time. I typically use this approach and it is, by far, the most commonly used approach among researchers.

*Hierarchical***.** Based on an *a priori* criteria, the researcher enters some number of variables into the model a step at a time. Any number of variables can be entered on each step, and any number of steps can be used. Each step is a separate regression model. The resulting model is identical to a model in which all variables were entered simultaneously. The major advantage of this approach is that a change in R-squared is computed, allowing for a test of whether a significant amount of additional variance is accounted for by the variable or variables entered on each step. If a single variable is entered on a step, the R-squared is equal to the semi-partial (a.k.a. "part") correlation coefficient, and the test of the R-squared change is equivalent to the test of the regression coefficient for the new variable.

Data Determine Model

Forward selection. Predictor variables are added to the model a step at a time. The first step evaluates all of the variables, and the variable with the largest correlation with the dependent variable is entered first. Then on each new step, the variable which will increase R-squared the most will be entered on that step (other criteria for particular significance levels—termed "PIN" for the p-value needed to be entered—or F values can be used). This approach is rarely used anymore.

Backward selection. Backward selection proceeds in the opposite manner to forward selection. All variables are entered and then the poorest predictor is eliminated. The process continues until all of the nonsignificant variables are removed. Usually by default variables that are not significant are removed on each step ("POUT" of .05), but any p-value or F-value can be used for the criteria. The model is reevaluated after each variable is removed. This approach is rarely used anymore.

Stepwise selection. Stepwise, which uses a combination of forward and backward selection, is more commonly used than either forward or backward. Predictor variables are entered as they are in forward selection, but at each step the variables are evaluated to see if any can be removed. As with the others, the criteria can be changed to a particular PIN and POUT or FIN and FOUT values.

All subsets and best subsets regression. All subsets regression (and closely related "best subsets regression" in SPSS) picks the best combination of predictors by running regression analyses for all possible predictors (according to the list provided). That is, if five predictors are given, there will be one 5-predictor model, five 4-predictor models and so on ($2^5 = 32$ models). One difficulty is deciding the optimal criteria to use in choosing the "best" model. Researchers may use a variety of criteria for picking the best possible model, including the highest R-squared, the lowest MSE (mean-squared residual or $MS_{residual}$), Akaike's Information Criterion (AIC; Akaike, 1973; or AICC, which is corrected for small samples; Hurvich & Tsai, 1995), Bayesian Information Criterion (BIC; Schwarz, 1978) or the lowest Mallows' C_p (Mallows, 1973). C_p is based on MSE but takes the number of predictors into account (models with more predictors always have higher R-squared values regardless of how useful the variables really are). In other contexts, you may see these indices are stated in terms of the likelihood function. Below are some of the formulas stated in terms of sum of squares residual (error), the standard error of the estimate, $sd_{Y-\hat{Y}}$, and p, which is the number of predictors plus the intercept.

$$R^2 = 1 - \frac{SS_{residual}}{SS_{total}} \quad AIC = n\log\left(\frac{SS_{residual}}{n}\right) + 2p + n + 2 \quad AICC = n\log\left(\frac{SS_{residual}}{n}\right) + \frac{n(n+p)}{n-p-2}$$

$$BIC = n\log\left(\frac{SS_{residual}}{n}\right) + 2(p+2)\left(\frac{n \cdot sd_Y^2}{SS_{residual}}\right) - 2\left(\frac{n \cdot sd_{Y-\hat{Y}}^2}{SS_{residual}}\right)^2 \quad C_p = \frac{SS_{residual}}{sd_{Y-\hat{Y}}^2} + 2p - n$$

The all subsets procedure is available in SAS with PROC REG and the best subsets procedure is available in SPSS under the LINEAR procedure. In *R*, you can use the `leaps` package or `regsubsets()` base function with the `method=c("exhaustive")`

Decision tree regression. Another highly popular data science or machine learning method is decision tree regression ("decision tree models" or sometimes just "decision trees"). There are a variety of specific decision tree approaches and algorithms, such as ID3 (iterative dichotomiser 3; Quinlan, 1986) is just for categorical data, C4.5 is an extension of the ID3 for continuous variables (Quinlan, 1993), and CART (classification and regression trees; Breiman et al., 1984). In essence, all are a kind of model selection method in which the most important predictor is selected out first (root node) and then additional predictors are selected in a branching fashion based on reduction of unaccounted for variance. There are too many concepts and terms, such as splits, edges, nodes, leafs, and Gini factors to cover here (see by Pathak et al., 2018 and James et al., 2023 for introductions), but the general approach fits into a purely exploratory framework that attempts to find the most important predictors for highly complex data sets or in the absence of *a priori* hypotheses. See Rachel Cabrera's review of R packages, https://rpubs.com/rachelcabrera/855412 and Simon Moss's introduction for SPSS https://www.cdu.edu.au/files/2020-07/Introduction%20to%20decision%20trees.docx.

Bagging, boosting, and random forests. Bagging and boosting are specific model selection algorithmic features used along with decision-tree modeling and commonly used in a few areas such as neural networks (see Sutton, 2005 for a review). Bagging or bootstrap aggregating (Breiman, 1996) is a resampling method based on subsets of cases in the data set. Multiple samples are combined to reduce error in prediction, but each sample is combined with equal weight to decide on inclusion of variables in the model (i.e., "voting"). Boosting (Freund & Schapire, 1996) also combines results from many analyses, but cases are weighted based on how well they are predicted and analyses are recomputed in an iterative process. Bagging and boosting are primarily used in engineering and computer science and relatively rarely used in the social sciences. Some evidence suggests that they can improve prediction, with some papers suggesting boosting outperforms bagging (Quinlan, 2006). Random forests (Ho, 1998) improves upon bagging using a random process of selection from a full set of predictors at each branching.

Relative weight analysis. Although not generally used as a model selection strategy per se, relative weight analysis (and several closely related approaches, such as dominance analysis and commonality coefficients) attempt to identify the relative importance of predictors in the model. Usually these approaches are not used for eliminating unimportant predictors or adding important predictors, but they are intended to find the most valuable predictors of some phenomenon. This is an admirable goal, as researchers need to move beyond just significance tests, but the several approaches that have been proposed all have some pitfalls and can often involve complicated computations (see Nimon & Oswald, 2013 for approaches in R, and see Tonidandel & LeBreton, 2011 for links to several macros). At the heart of most relative weight type measures are standardized coefficients or squared multiple correlations (for an introduction, see Tonidandel & LeBreton, 2011). The relative weight methods try to improve upon squared standardized coefficients, which are subject to inaccuracies in the ability to sum to the total R-squared value whenever the predictors are correlated, as measures of proportion of variance accounted for uniquely by a predictor. Often the various approaches arrive at similar results and are correlated with one another (Johnson, 2000). I am not certain how much they add beyond an increment in R-squared (or semi-partial correlation) coefficient and it must be kept in mind that R-square change and any of these measures are subject to sampling variability, sometimes with very wide intervals. This makes replication critical to determine real relative importance.

Comments

Hierarchical regression can be useful for estimating the variance accounted for by sets of variables. But one reason I usually use simultaneous regression rather than hierarchical regression is that all coefficients are partial with respect to all other variables considered. Researchers who use hierarchical regression might enter demographic variables on the first step as "control variables" or "covariates." There are no substantial differences in the two approaches, however. Hierarchical regression is simply running several different regression models at the same time, each one with added predictors. The one

difficulty I have with hierarchical regression merely involves the usual format of presentation. Because researchers often present results for only the new variables entered on a particular step, readers cannot tell what happens to variables entered on prior steps after new variables have been entered. For example, if age becomes nonsignficant after another variable is entered on the second step, the reader will conclude that age was an important predictor even though a variable entered later was responsible for the association of age with the dependent variable. Your textbook (Cohen et al., 2003, Chapter 5, Section 5.3) makes a case for the order of entry with hierarchical regression based on causal ordering of variables. In general, I think hierarchical regression provides little advantage in understanding causal ordering and that causality should be explored in other ways (more detail when I discuss longitudinal analysis and mediation).

Forward and backward procedures are rarely used anymore, because stepwise selection is considered superior to either. Although stepwise selection is better than forward or backward alone, it still has problems. Simulation studies suggest that stepwise selection often leads to erroneous model choices (both Type 1 and Type 2 errors can occur; Freedman, 1983; Pope & Webster, 1972). I recommend that researchers use theory to decide which predictors to include and which order to enter variables into the model, rather than use exploratory, data-driven approaches. If a researcher wishes to go completely exploratory, the all subsets, boosting, and random forest decision trees may be preferred over the other exploratory approaches.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Pet rov & F. Caski (Eds.), Proceedings of the Second International Symposium on Information Theory (pp. 267-281). Budapest: Akademiai Kiado.

Breiman, L., Friedman, J., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees.* Chapman & Hall. New York.

Breiman, L. (1996a). Bagging predictors. *Machine Learning, 24*, 123–140.

Freedman, D.A. (1983) A note on screening regression equations. *The American Statistician, 37*, 152-155.

Freund, Y., Schapire, R. (1996). Experiments with a new boosting algorithm. In: Saitta, L. (Ed.), Machine *Learning: Proceedings of the Thirteenth International Conference*. Morgan Kaufmann, San Francisco,CA.

Ho TK (1998). "The Random Subspace Method for Constructing Decision Forests" (PDF). *IEEE Transactions on Pattern Analysis and Machine Intelligence. 20* (8): 832–844. doi:10.1109/34.709601

Hurvich, C. M., & Tsai, C.-L. (1995). Model selection for extended quasi-likelihoodmodels in small samples. Biometrics, 51, 1077-1084.

James, G., Witten, D., Hastie, T., Tibshirani, R., & Taylor, J. (2023). Tree-based methods. In *An introduction to statistical learning with applications in Python* (pp. 331-366). Springer International Publishing. https://datamineaz.org/readings/ISL_chp8.1.pdf

Johnson, J. W. (2000). A heuristic method for estimating the relative weight of predictor variables in multiple regression. *Multivariate behavioral research, 35*(1), 1-19.

Mallows, C. L. (1973). Some Comments on Cp. *Technometrics, 15*, 661–675.

Nimon, K. F., & Oswald, F. L. (2013). Understanding the results of multiple linear regression: Beyond standardized regression coefficients. *Organizational Research Methods, 16*(4), 650-674.

Pathak, S., Mishra, I., & Swetapadma, A. (2018, November). An assessment of decision tree based classification and regression algorithms. In 2018 3rd International Conference on Inventive Computation Technologies (ICICT) (pp. 92-95). *IEEE Proceedings of the International Conference on Inventive Computation Technologies* (ICICT-2018).

Pope, P.T., & Webster, J.T. (1972). The use of an F-statistic in stepwise regression procedures. *Technometrics, 14*, 327-340.

Quinlan, J. R. (1986). Induction of decision trees. *Machine learning, 1*, 81-106.

Quinlan, J. R. (1993). *Programs for Machine Learning C4. 5*. Morgan Kaufmann Publishers, San Francisco, CA.

Quinlan, J. R. (2006). Bagging, Boosting, and C4.5. In *AAAI 96: Proceedings of the 13th National. Conference on Artificial Intelligence* (Vol. 2, pp. 725–730).

Tonidandel, S., & LeBreton, J. M. (2011). Relative importance analysis: A useful supplement to regression analysis. Journal of Business and Psychology, 26(1), 1-9.

Schwarz, G. (1978). Estimating the dimension of a model. Annals of Statistics, 6, 461-464

Sutton, C.D. (2005). Classification and Regression Trees, Bagging, and Boosting, in *Handbook of Statistics*, Vol. 24, pp. 303-329, Elsevier.