Variable Selection

Variable selection differs, depending on the goal of the model.

Goal of the Model

Researchers usually have more potential predictor variables than end up in the final model. What variables to include is largely a question of the goal of the model. There are usually three possible goals:

1. *Prediction.* Here the primary issue is minimizing prediction error rather than causal interpretation of the predictors in the model….

2. *Evaluating a predictor of primary interest.* In pursuing this inferential goal, a central problem in observational data is confounding, which relatively inclusive models are more likely to minimize. Predictors necessary for *face validity* as well as those that behave like confounders should be included in the model….

3. *Identifying the important independent predictors of an outcome*. This is the most difficult of the three inferential goals, and one in which both causal interpretation and statistical inference are most problematic. Pitfalls include false-positive associations, the protential complexity of causal pathways, and the difficulty of identifying a single best model. We also endorse inclusive models in this context, and recommend a selection procedure that affords increased protection against false-positive results. Cautious interpretation of weak associations is key to this approach."

Evaluating a predictor of primary interest

If the goal is to evaluate a predictor of primary interest, then eliminating variables based solely on statistical significance is not the best approach.

Vittinghoff et al (2005, p.146) state,

> "However, we do not recommend 'parsimoniuous' models that only include predictors that are statistically significant at $P < 0.05$ or even stricter criteria, because the potential for residual confounding in such models is substantial."

Maldonado and Greenland (1993) suggest that potential confounders be eliminated only if $p > 0.20$, in order to protect against residual confounding.

For the other goals of identifying the list of important predictors or to develop a prediction model, than retaining only significant predictors makes sense.

"10% change in estimate" variable selection rule

Confounding is said to be present if the unadjusted effect differs from the effect adjusted for putative confounders. (Rothman, 1998).

A variable selection rule consistent with this definition of confounding is the *change-in-estimate* method of variable selection. In this method, a potential confounder is included in the model if it changes the coefficient, or effect estimate, of the primary exposure variable by 10%. This method has been shown to produce more reliable models than variable selection methods based on statistical significance (Greenland, 1989).

Protocol Suggestion

Grant reviewers like to see some discussion about variable selection. Here is some suggested wording for the change-in-estimate method.

> Given that the goal of the multivariable model is to assess the effect of the study intervention, while controlling for putative confounding variables, variable selection will be done using the *change-in-estimate* method. This method has been shown to produce more reliable models than variable selection methods based on statistical significance (Greenland, 1989). In this method, a potential confounder is included in the model if it changes the coefficient of the primary exposure variable, our study intervention, by 10 percent. This approach is consistent with the definition of confounding, where confounding is said to be present if the unadjusted effect differs from the effect adjusted for putative confounders (Rothman and Greenland, 1998).

Using Both 10% Rule and P Values

The most common approach is to just use significance for variable selection. It is rare to see just the 10% rule being used, even though it is a better approach. Frequently authors use a combination of the the two approaches.

This is consistent with what was said on page 1 under the heading "2. Evaluating a predictor of primary interest", where it was mentioned that variables to provide face validity be included.

Example 1) "10% change in estimate" and statistical significance variable selection

Kulkarni et al (*N Engl J Med*, 2006) state in their Statistical Analysis section,

> "In the multiple regression models, confounders were included if they were significant at a 0.05 level or they altered the coefficient of the main variable by more than 10 percent in cases in which the main association was significant."

Example 2) "10% change in estimate" and statistical significance variable selection

Chaves et al (*N Engl J Med*, 2007) state in their Statistical Analysis section,

> "We examined any association between potential predictors of increased severity of disease spearately for subjects who were vaccinated and those who were not vaccinated, using a two-sided chi-square test. We constructed two unconditional logistic-regression models—one for vaccinated subjects and one for unvaccinated subjects—to determine which variables remained independent predictors that subjects would have moderate-to-severe disease. Variables that had a significant association with disease severity in the univariate analysis were included in the multivaraite regression models. Variables that were not significantly associated with disease severity but that changed the odds ratio for severity by 10% or more when removed from the analysis were also kept in the final model. (Maldonato et al.)"

More Cautious Approach to Guard Against Confounding (10% Rule + conservative P value + a priori confounders)

Since confounding does not depend on statistical significance, nor is the 10% rule a definitive cutpoint for defining confounding, some investigators take a more cautious approach. Thompson et al (*N Engl J Med*, 2007) provides a good example in their Statistical Analysis section,

> "We analyzed raw test scores adjusted for a priori confounders, including linear terms for age, family income, and score on the HOME scale[14,15] and dummy-coded variables for sex, HMO, maternal IQ, maternal education, single-parent status, and birth weight. Other covariates were included in the full model if the P value was less than 0.20 or if their inclusion resulted in a change of 10% or more in the estimate of the main effect of mercury exposure (Maldonado et al., Budtz-Jørgensen et al.…)"

Backwards Elimination

Backwards selection is considered superior to forwards selection (forward selection adds one variable at a time), because negatively confounded sets of variables are less likely to be omitted from the model (Sun et al, 1999), since the complete set is included in the initial model. In contrast, forward and stepwise (stepwise is where variables can be added and subsequently removed) selection procedures will only include such sets if at least one member meets the inclusion criterion in the absence of the others. (Vittinghoff et al, 2005, p.151).

By "negatively confounded sets" , we are referring to the situation where two or more variables must be included in the model as a set to control for confounding. When one of the variables is dropped, confounding increases.

Budtz-Jørgensen et al (2006) recommend using p=0.20 as the cut-off when backwards elimination is used. In the Tompson et al (2007) example shown above on this page, the researchers use p=0.20 and cite Budtz-Jørgensen.

Automated Variable Selection Procedures

Statistical software packages provide automated variable selection routines, giving you the choice of forward, backward, or stepwise. Although these were once popular, they have fallen under enough criticism that it is very rare to find an article that admits to using them. These automated routines, although finding a significant set of predictors, have no way to make decisions about collinearity or confounding, and they can even produce nonsensical models (Greenland, 1989).

A better approach, then, is to use "interactive backwards elimination", where you, the researcher, makes the decision at each step.
The automated variable selection routines are available in R for any type of regression model.

References

Budtz-Jørgensen E, Keiding N, Grandjean P, Weihe P. (2006). Confounder selection in environmental epidemiology: assessment of health effects of prenatal mercury exposure. *Ann Epidemiol* 17:27-35.

Chatterjee S, Hadi AS, Price B. (2000). Regression Analysis by Example. 3rd ed. New York, John Wiley and Sons.

Chaves SS, Gargiullo P, Zhang JX. (2007). Loss of vaccine-induced immunity to varicella over time. *N Engl J Med* 356(11):1121-9.

Dupont WD. (2002). *Statistical Modeling for Biomedical Researchers: A Simple Introduction to the Analysis of Complex Data*. Cambridge UK, Cambridge University Press.

Greenland S. (1989). Modeling and variable selection in epidemiologic analysis. *Am J Public Health* 79(3):340-349.

Kulkarni N, Pierse N, Rushton L, Grigg J. (2006). Carbon in airway macrophages and lung function in children. N Engl J Med 355(1):21-30.

Maldonado G, Greenland S. (1993). Simulation study of confounder-selection strategies. *Am J Epidemiol* 138:923-936.

Rothman KJ, Greenland S. (1998). *Modern Epidemiology*, 2nd ed. Philadelphia, PA, Lippincott-Raven Publishers.

Steyerberg EW. (2009). *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*. New York, Springer.

Thompson WW, Price C, Goodson B, et al. (2007). Early thimerosal exposure and neuropsychological outcomes at 7 to 10 years. *N Engl J Med* 357;13:1281-1292.

Vittinghoff E, Glidden DV, Shiboski SC, McCulloch CE. (2005). *Regression Methods in Biostatistics: Linear, Logistic, Survival, and Repeated Measures Models*. New York, Springer.