ERHS 642 Logistic Regression Spring 2016

In-class Assignment 8

Austin et al., Automated variable selection methods for logistic regression

1. What are possible reasons for the different sets of model covariates in studies of the same exposure and outcome of interest?

2. On page 1139, 1st full paragraph, the authors state that "investigators developing models to predict mortality need to maintain a balance between including too many variables and model parsimony". Explain why.

3. Explain the difference between forward, backward and stepwise selection.

4. Austin and Tu only included variables univariately significant at the 0.25 level in the automated selection processes. What are the pros and cons of this approach?

5. What is a bootstrap sample and how was it used in the study by Austin and Tu?

6. Is model building based on clinical judgment in combination with an automated process always an improvement over using an automated process only?

7. The variable “respiratory rate” was chosen in approximately 375 of the 1000 models (page 1142, figure 1); the variable “heart rate” was chosen in almost 900 of the 1000 models.   Do you think either variable is a true predictor of mortality after AMI?

8. Are the results shown in figure 2 surprising?

9. Explain how noise variables were created, how they were used and summarize the results for noise variables.

10. In light of noise variables, why should the number of model covariates be kept as small as possible?

11. What are the pros and cons of automated model building for explanatory, predictive and exploratory models?

12. Suggest improved approaches to model building.

13. Summarize the article’s main conclusions.