The Journal of Foot & Ankle Surgery 58 (2019) 403–404

Contents lists available at ScienceDirect

The Journal of Foot & Ankle Surgery

journal homepage: www.jfas.org

Investigator's Corner

Model Selection: Finding the Right Fit

Daniel C. Jupiter, PhD

Associate Professor, Department of Preventive Medicine and Community Health, The University of Texas Medical Branch, Galveston, TX

ARTICLE INFO

Keywords:
cross-validation
descriptive models
predictive modeling
risk factor
stepwise regression

ABSTRACT

There are numerous possible goals for building statistical models. Those statistical goals, the associated model types, and each statistical tool involved in model building come with its own assumptions and requirements. In turn, these requirements must be met if we are to ensure that our models produce meaningful, interpretable results. However, beyond these technical details is the intuition, and the additional set of tools and algorithms, used by the statistician, to build the contextually appropriate model: not only must we build an interpretable model, we must build a model that answers the particular question at hand and addresses the particular goal we have in mind. In this column we discuss the methods by which statisticians build models for description, risk factor identification, and prediction.

In this commentary, our last and shortest on the topic of the building of statistical models, we in a sense come full circle. We have discussed the nitty gritty technical details of model assumptions (1). We have gotten into the subtle differences between different model types or, more properly, between the various goals of modeling and the particulars of model quality and goodness of fit (2). We have delved into the various specialized tools used for assessing whether we have met our assumptions and the quality of our models (3). In our most recent commentary, we examined why making predictions is so difficult a task for statisticians and how that challenge is met by using some novel tools (4). In this commentary, we step back a bit, leaving behind some of the technicalities, and discuss, slightly more generally, how statisticians put all the pieces together, when they are actually involved in the business of building models. Reality, the world around us, the medical world we are tasked with exploring, is creeping back into the picture; we leave behind the inner workings of the statistical machinery and try to ensure that we are answering the question for which our study was designed in the first place.

Given all the tools we have, given all the assumptions required of our statistical models, given the technicalities of making predictions, or identifying risk factors, or building descriptions of the world ... how does a statistician, pondering a clinical project, go about deciding which model to build? More specifically, how should he or she choose which variables should be included in a given model? How should the statistician then go about developing, tweaking, and finalizing a model? There are many approaches, and opinions may differ on how these tasks are best to be handled. Here we give some ideas, some suggestions of how the various models can be built. These are opinions, however, and the statisticians with whom you work may have different, and equally valid, methodologies. Our goal, rather than being entirely prescriptive, or describing exactly how tasks are to be done in each possible setting, is to give an overview of the process and to remind the reader, again, that not all statistical modeling tasks are created equal.

Real World Considerations

Regardless of whether our goal is building a description of how the world works, identifying predisposing factors, or making predictions, we are in great part trying to capture a notion of reality. We are trying to either inform our intuition by capturing something about the world in our model or, conversely, trying capture something about our intuition in a model. In all cases, our prior knowledge of reality should guide our statistical efforts. If we know that a particular variable impacts our outcome of interest or that the expectation in the field in which we are working is that that variable is accounted for, we should include it in our model. This is regardless of any issues of parsimony, fidelity, or significance.

Descriptive Modeling

As we recall from an earlier commentary (2), we sometimes have in mind the goal of understanding what our world looks like: we are not here trying to identify risk factors, nor are we making predictions about individuals. We simply want a view of the "machine" as a whole; how does it run? As an example, let us consider, as we have before, patients who need a repair for a particular type of fracture, and let us aim to explore time to healing. Screws and plates are our methods of fixation, and we can consider as well many other factors that go into management of fractures, patient health, and length of time until union: gender,

Address correspondence to: Daniel Jupiter, PhD, 301 University Boulevard, 1.134G Ewing Hall, Galveston, TX 77555-1150.
E-mail address: dajupite@utmb.edu

1067-2516/$ - see front matter © 2019 by the American College of Foot and Ankle Surgeons. All rights reserved.
https://doi.org/10.1053/j.jfas.2019.01.009

race, age, diabetes status, severity of injury, open versus closed fracture, isolated versus nonisolated injury, facility type, rehabilitation protocol, etc. It is not possible to include all of these variables in our model. It may not even be possible, depending on the data source, to obtain information regarding all the variables. Our goal is to obtain a parsimonious, relatively accurate picture of how patients fare, on average. What does the patient mix look like? How do different groups of patients progress through treatment and recovery?

The accuracy we desire in our model is balanced against the efficiency and parsimony we would also like to achieve in that same model. Further, statistical significance of our covariates is not a requisite. We recall that it is often the case that the accuracy of descriptive models, in the model building process, will be assessed with tools such as the Akaike information criterion (AIC) or Bayesian information criterion (2). These balance the need for an efficient model, against the gains in precision that can be obtained by including more variables in our models.

We recall that addition of new variables to a model always improves its fit, in terms of R^2 (2). How then do we decide which variables to include in our model? How can we use tools such as the AIC, adjusted R^2, or others to balance parsimony against accuracy? One commonly used algorithm for this decision making process is called forward selection. The idea with forward selection is that given a particular model of fracture healing, as above, we could consider adding any one of the remaining variables that are not yet included in the model. We choose to add that single unique variable with smallest p value, with the proviso that that p value is below a prespecified cutoff. We then repeat the process, adding variables one at a time, selecting that with smallest p value on inclusion in the model, until no variable meets criteria to be added the model. Generally, this process is started from a model with no covariates included, or just those which we demand are included for real world reasons. The final model obtained by this process, regardless of the statistical significance of any covariate included in it, will serve as our model of the world.

An analogous procedure, called backwards selection, starts with all variables included in the model and removes those in turn that are least significant and with p value above a prespecified threshold. A third procedure, stepwise regression, combines the 2 processes, adding and subtracting variables in a search for a good model. These procedures can be adapted to examine criteria other than p values, such as the adjusted R^2, AIC, or Bayesian information criterion. In all cases, we are searching for a model that is in some way optimal. Stepwise procedures do this in a methodical, though not necessarily entirely efficient or exhaustive, way. There are additional algorithms to do deeper searches, of all possible models, that may be more inclusive; we do not go into detail here. However, we do remark that this exhaustive searching leads to 2 criticisms. First, with all the statistical testing we are doing, we have a multiple test correction problem (5). Second, we may be overfitting our model, with the model being a good fit for our particular data set but not generalizable. These caveats should be kept in mind and point to the fact that this selection procedure should not be carelessly used for identification of risk factors, or discovering predictive models.

Identification of Risk Factors

In thinking about risk factors, we are generally looking for variables that have a significant association with the outcome of interest. Further, we are usually looking for independent risk factors, as described in earlier commentaries (2), and especially in the context of comparing and contrasting with predictive models (4). That is, we are specifically interested in those variables with a significant relationship with the outcome, in a multivariable model.

How do we choose these variables, and these models, while avoiding the overfitting and multiple test correction issues encountered in stepwise regression? One method is relatively simple ... and intuitive. We perform the bivariate analyses needed to assess each covariate's relationship to the outcome variable. We then choose to include in our final model all those variables whose bivariate test yielded a p value below some chosen cutoff; that cutoff should be above our standard .05, perhaps .1 or .2. The intuition here is that we are looking for variables that could possibly have some sort of relationship with the outcome in multivariate analysis: this eliminates, or most likely eliminates, variables with p values close to 1. On the other hand, those variables with a bivariate relationship to the outcome whose p value is close to significance should be included in the multivariate model: until we put them in the multivariable model, we do not know how interrelationships between independent variables might impact the relationship of each with the outcome variable. Our final model, including these candidate variables, is then used to define the risk factors: those variables with significant association with the outcome, in the multivariate model, are risk factors.

It is worth noting that this is not the only method for identification of risk factors. Indeed, some statisticians will use tools similar to the stepwise regression, or the searches for optimal models, mentioned earlier, to identify risk factors. This can be done despite our warnings! Our caveats presented above are merely there to warn us that, in unskilled hands, searches for optimal models may lead to issues of overfitting.

Making Predictions

Our last commentary dealt in depth with the nature of predictive models, so little more needs to be said here (4). As compared with building descriptive models, the goal with predictive models is to ensure that, regardless of significance or measures such as the AIC, the predictions of the model on future subjects are accurate. In contrast with models aimed at discovering risk factors, the significance of individual covariates is not as important, and variables are not singled out for specific mention. However, the search for the most appropriate model remains important and, as described in our last commentary, the results of that search are often described using something such as cross-validation. Usually, those results are not described with p values but rather by quantifying the accuracy of predictions and comparing such accuracy between candidate models. While overfitting is still a concern, at least partially addressed by cross-validation, multiple test correction errors are not.

Our circuit is now complete. We have looked at why the assumptions of statistical models are important and how models are assessed. We have examined tools for testing that our assumptions are met, and we have examined predictive models in detail. Finally, we have thought a little about how statisticians approach building different types of models. The story is far from over, and much more can be said about each of these topics. But the hope is that this overview has provided some insight into the model building and selection process that you can carry into conversations with your statistician collaborators.

References

1. Jupiter D. Investigators' Corner: Assumptions of statistical tests: what lies beneath. J Foot Ankle Surg 56:910–913, 2017.
2. Jupiter D. Investigators' Corner: Snug as a Bug: Goodness of Fit and Quality of Models. J Foot Ankle Surg 56:1357–1360, 2017.
3. Jupiter D. Investigators' Corner: The doctor is in! Diagnostic Analysis. J Foot Ankle Surg 57:427–431, 2018.
4. Jupiter D. Investigators' Corner: Seeing the future: a crystatistical ball. J Foot Ankle Surg 57:850–853, 2018.
5. Jupiter D. Investigators' Corner: Multiple choice answers: what to do when you have too many questions. J Foot Ankle Surg 54:285–286, 2015.