STATISTICS IN MEDICINE, VOL. 15, 361–387 (1996)

TUTORIAL IN BIOSTATISTICS

MULTIVARIABLE PROGNOSTIC MODELS: ISSUES IN DEVELOPING MODELS, EVALUATING ASSUMPTIONS AND ADEQUACY, AND MEASURING AND REDUCING ERRORS

FRANK E. HARRELL Jr., KERRY L. LEE AND DANIEL B. MARK

Divisions of Biometry and Cardiology, Box 3363, Duke University Medical Center, Durham, North Carolina 27710, U.S.A.

SUMMARY

Multivariable regression models are powerful tools that are used frequently in studies of clinical outcomes. These models can use a mixture of categorical and continuous variables and can handle partially observed (censored) responses. However, uncritical application of modelling techniques can result in models that poorly fit the dataset at hand, or, even more likely, inaccurately predict outcomes on new subjects. One must know how to measure qualities of a model's fit in order to avoid poorly fitted or overfitted models. Measurement of predictive accuracy can be difficult for survival time data in the presence of censoring. We discuss an easily interpretable index of predictive discrimination as well as methods for assessing calibration of predicted survival probabilities. Both types of predictive accuracy should be unbiasedly validated using bootstrapping or cross-validation, before using predictions in a new data series. We discuss some of the hazards of poorly fitted and overfitted regression models and present one modelling strategy that avoids many of the problems discussed. The methods described are applicable to all regression models, but are particularly needed for binary, ordinal, and time-to-event outcomes. Methods are illustrated with a survival analysis in prostate cancer using Cox regression.

1. INTRODUCTION

Accurate estimation of patient prognosis is important for many reasons. First, prognostic estimates can be used to inform the patient about likely outcomes of her disease. Second, the physician can use estimates of prognosis as a guide for ordering additional tests and selecting appropriate therapies. Third, prognostic assessments are useful in the evaluation of technologies; prognostic estimates derived both with and without using the results of a given test can be compared to measure the incremental prognostic information provided by that test over what is provided by prior information.[1] Fourth, a researcher may want to estimate the effect of a single factor (for example, treatment given) on prognosis in an observational study in which many uncontrolled confounding factors are also measured. Here the simultaneous effects of the uncontrolled variables must be controlled (held constant mathematically if using a regression model) so that the effect of the factor of interest can be more purely estimated. An analysis of how variables (especially continuous ones) affect the patient outcomes of interest is necessary to

CCC 0277–6715/96/040361–27

ascertain how to control their effects. Fifth, prognostic estimation is useful in designing randomized clinical trials. Both the decision concerning which patients to randomize and the design of the randomization process (for example, stratified randomization using prognostic factors) are aided by the availability of accurate prognostic estimates before randomization.[2] Lastly, accurate prognostic models can be used to test for differential therapeutic benefit or to estimate the clinical benefit for an individual patient in a clinical trial, taking into account the fact that low-risk patients must have less absolute benefit (lower change in survival probability).[3]

To accomplish these objectives, analysts must create prognostic models that accurately reflect the patterns existing in the underlying data and that are valid when applied to comparable data in other settings or institutions. Models may be inaccurate due to violation of assumptions, omission of important predictors, high frequency of missing data and/or improper imputation methods, and especially with small datasets, overfitting. The purpose of this paper is to review methods for examining lack of fit and detection of overfitting of models and to suggest guidelines for maximizing model accuracy. Section 2 covers initial steps such as imputation of missing data, pre-specification of interactions, and choosing the outcome model. Section 3 has an overview of the need for data reduction. In Section 4, we discuss the process of checking whether a hypothesized model fits the data. In Section 5, measures of predictive accuracy are covered. These are not directly related to lack of fit but rather to the ability of the model to discriminate and be well calibrated when applied prospectively. Section 6 covers model validation and demonstrates advantages of resampling techniques. Section 7 provides one modelling strategy that takes into account ideas from earlier sections and lists some miscellaneous concerns. Most of the methods presented here can be used with any regression model. Section 8 briefly describes some statistical software useful in carrying out the strategy summarized in Section 7. Section 9 has a detailed case study using a Cox regression model for time until death in a clinical trial studying prostate cancer.

2. PRELIMINARY STEPS

Before analyses begin, the researcher must specify the relationships of interest and define and assemble the response variable and the potential predictors. At this point a frequent problem is the extent of missing data. Some methods of dealing with missing data are given in References 4–7. Deletion of cases with missing predictors causes bias and increased variance. Even though caution should be taken when imputing missing values, it is usually better to estimate selected data values than to delete an entire subject's record. Simple methods of imputation include the use of the median, mean, or mode for missing values. This method is biased and inefficient when predictors are correlated with one another.[4] Deriving customized regression models for predicting each predictor from all other predictors is a better method. Kuhfeld[8] has implemented a general imputation method that allows predictors to be non-linearly (and even non–monotonically) related to one another. This method has been modified by Harrell and implemented in the S-Plus `transcan` function (Section 8), which yields stable imputations even when the fraction of missing values is quite large. In some cases, surrogate predictors, not intended to enter the model directly, are assembled to assist in imputing missing predictors in the model.

It is important that maximum information be extracted from predictors and response. Because of this and because of problems with data reliability, when one has a choice of describing a concept with a categorical variable or a continuous one, the continuous one is preferred. Subject matter knowledge should guide the selection of candidate predictors. Early deletion of those with little chance of being predictive or of being measured reliably will result in models with less overfitting and greater generalizability.

Plausible interactions should be carefully chosen because of problems of multiple parameters (see reference 9 for additional thoughts on interactions). Certain types of interactions that have frequently been found to be important in predicting clinical outcomes and thus may be pre-specified are:

1. Interactions between treatment and the severity of disease being treated. Patients with little disease have little opportunity to receive benefit.
2. Interactions involving age and risk factors. Old subjects are generally less affected by risk factors. They have been robust enough to survive to their current age with risk factors present.
3. Interactions involving age and type of disease. Some diseases are incurable and have the same prognosis regardless of age. Others are treatable or have less effect on younger patients.
4. Interactions between a measurement and the state of a subject during a measurement. For example, left ventricular function measured at rest may have less predictive value and thus have a smaller slope versus outcome than function measured during stress.
5. Interactions between calendar time and treatment. Some treatments evolve or their effectiveness improves with staff training.
6. Interactions between quality and quantity of a symptom.

Careful fitting of a statistical model is essential so that interactions, if present, represent biologic phenomena rather than general lack of fit of the model.

A tentative choice of the statistical model is sometimes based on previous distributional examinations, but it is frequently based on maximizing how available information is used. Binary and ordinal logistic models[10–13] are frequently used for discrete completely assessed outcomes, and the Cox proportional hazards model[14,15] and parametric survival models[16] are frequently used for censored time-to-event data. It is quite common to change the model after initial modelling of predictors, because only then can adjusted distributional properties of Y and joint properties of X and Y be assessed (Section 4.3).

3. DATA REDUCTION

Multivariable statistical models when developed carefully are excellent tools for making prognostic predictions. However, when the assumptions of a model are grossly violated or when a model is used unwisely for a given patient sample, the performance of the model may be poor. For example, when the analyst has fitted not only real trends that further data would support, but in addition has fitted idiosyncrasies in the particular dataset by analysing too many variables, the model may predict inaccurately for a new group of patients. Only with appropriate model validation can an apparently accurate model be shown to be inaccurate.

In developing a set of predictions based on 100 patients, no analyst would divide the patients into 50 subgroups and quote the average outcome for each subgroup. Yet many articles have appeared in the clinical literature where 20–50 variables were analysed on 100 patients. Researchers apparently do not realize that when many predictor variables are analysed, variable screening based on statistical significance and stepwise variable selection involve multiple comparisons problems that lead to unreliable models. These methods are therefore not viable for data reduction (see Reference 17 for a condemnation of stepwise variable selection).

The situation is actually worse than merely considering the number of predictors. If the analyst used associations with Y to entertain non-linearities in the predictors or interaction terms, these constructed variables need to be counted (see Table II for an example). We speak of the total

predictor degrees of freedom (d.f.), p, as the total number of parameters (columns of the design matrix) examined during the course of analysis, excluding intercept term(s). If graphical or other informal analyses are used to guide the analysis, it is difficult to define p – one needs to estimate the effective number of parameters considered according to the flexibility of fits that were considered.[18] The quantity p is the effective number of parameters allowed for consideration, that is, the number of regression coefficients estimated formally or informally without algebraic restrictions.

To enhance the accuracy of a model, the number of variables used must be reduced or the model must be simplified unless the sample is large. Unless a formal penalized estimation technique is used,[19] multiple comparisons problems that arise from 'peeking' at the outcome variable must be eliminated; data reduction methods must be used that do not utilize the outcome variable. Harrell *et al.*[20] discussed some available data reduction methods and two regression modelling strategies based on these methods that yield reliable models. They suggest as a rough rule of thumb that in order to have predictive discrimination that validates on a new sample, no more than $m/10$ predictor d.f. p should be examined to fit a multiple regression model, where m is the number of uncensored event times (for example, deaths) in the training sample (the sample used in fitting the model). For binary outcomes m is the number of patients in the less frequent outcome category. If $p > m/10$, a data reduction technique such as principal components, variable clustering, or deriving clinical summary indexes[20–23] should be used until the number of summary variables to use as candidates in the regression analysis is less than $m/10$.

Smith *et al.*[24] found in one series of simulations that the expected error* in Cox model predicted 5-year survival probabilities was below 0·05 when $p < m/20$ for 'average' subjects and below 0·10 when $p < m/20$ for 'sick' subjects. For 'average' subjects, $m/10$ was adequate for preventing expected errors $> 0{\cdot}1$.

Better and more general than any of these rules is the reduction of d.f. using a shrinkage method (Section 5.4).

4. VERIFYING MODEL ASSUMPTIONS: CHECKING LACK OF FIT

4.1. Linearity assumption

In their simplest forms, all usual regression models assume that for a certain scale of Y, each predictor variable X is linearly related to Y. In the logistic regression model for binary responses, the initial assumption is that an X is linearly related to the log odds of response ($\log[P/(1-P)]$, where P is the probability of response) for patients subgrouped by values of X. In the Cox proportional hazards survival model, one initially assumes that at each time t, $\log[-\log(S(t))]$ and equivalently $\log\lambda(t)$ are linearly related to X, where $S(t)$ is the probability of surviving until time t and $\lambda(t)$ is the hazard function or instantaneous event rate at time t. It is easy to envision cases where strong violations in the linearity assumption (say a U-shaped age relationship) will result in erroneous predictions.

A direct way to check the linearity assumption, and to determine how to transform a specific X if necessary, involves expanding X into multiple terms that can flexibly fit any smooth relationship. The extra terms can be statistically tested to assess the adequacy of a linear relationship, and the terms *in toto* can estimate the true transformation of X that would result in

* Absolute difference between predicted and actual 5-year survival probabilities in a simulation study with known survival functions

a linear relationship with Y. A common choice of expansion is to add X^2 and perhaps higher powers of X to the model. A more flexible approach is the use of piecewise linear regression or piecewise cubic polynomials (spline functions). See references 25–27 for methods of fitting such functions.

As an alternative, smoothed residual plots can be used to determine the functional form for each predictor. For binary logistic models, smoothed partial residual plots[13,28,29] are useful, and for the Cox model, smoothed martingale residuals plots detect regression shape departures.[30] Partial residuals in logistic models are particularly computationally efficient, as the analyst can fit a simple model that is linear in all predictors and then use the residuals to obtain estimates of the *true* functional forms. However, the plot for each predictor does assume that the other predictors operate linearly and that all predictors are additive (see below). The usual martingale residual plot for the Cox model provides an estimate of the *departure* from linearity for the predictor.

4.2. Additivity assumption

A further assumption of most regression models is additivity of effects of the predictors (lack of interaction). Interactions can be tested and described by adding cross-product terms. It must be borne in mind that interactions can take the form of a change in shape (for example, linear age relationship for males, quadratic for females), so the cross-products needed in the model are not always simple ones.

The number of possible cross-product terms is usually so large (especially when variables have non-linear or multiple dummy variable components) that the predictors to check for additivity must usually be specified before examining the data. Otherwise, type I errors and overfitting will be significant problems. A compromise solution is to do pooled interaction tests. For example, in a model with predictors age, sex, and dose, one may test all second-order interactions involving age, all interactions involving sex, and all involving dose. A combined test of all two-way interactions is also useful. If a pooled test is not significant, it may be unwise to pursue significant component interactions.

4.3. Distributional assumption

The previous sections dealt with the proper specification of the X-structure of the model. Once the analyst has determined which predictors are to be used and how they should be represented in the model, most models have distributional assumptions that also need verification. The Cox model does not assume anything about the survival function $S(t)$ across t for an individual, but it does assume how survival curves for different subjects are related. Specifically, it assumes that $\log[-\log(S(t))]$ for different subjects are equidistant over time, or equivalently that hazard functions for any two subjects are proportional over time. This proportional hazards assumption can be checked using smoothed plots of a special type of residual from the model called the Schoenfeld residual.[31,32] It can also be checked using hazard ratio plots, plots of modelled versus stratified estimates,† and several other methods.[33] Unlike the Cox model, fully parametric models (for example, Weibull or log-normal survival models) have a distributional assumption even when there are no covariables. If the form of $S(t)$ does not fit the data for these models, estimates of $S(t)$ will be inaccurate.

† That is, a Cox model is fitted with the variable in question appearing as a covariate for which regression coefficient(s) are estimated, then a second model is fitted where that variable is used as a stratification factor that modifies the underlying survival function (but which does not have regression coefficients).

5. QUANTIFYING PREDICTIVE ACCURACY

There are at least three uses of measures of predictive accuracy:

1. To quantify the utility of a predictor or model to be used for prediction or for screening to identify subjects at increased risk of a disease or clinical outcome.‡
2. To check a given model for overfitting (fitting noise resulting in unstable regression coefficients) or lack of fit (improper model specification, omitted predictors, or underfitting). More will be said about this later.
3. To rank competing methods or competing models.

The measures discussed below may be applied to the assessment of a predictive model using the same sample on which the model was developed. However, this assessment is seldom of interest, as only the most serious lack of fit will make a model appear not to fit on the sample for which it was tailor-made. Of much greater value is the assessment of accuracy on a separate sample or a bias-corrected estimate of accuracy on the training sample. This assessment can detect gross lack of fit as well as overfitting, whereas the *apparent* accuracy from the original model development sample does not allow one to quantify overfitting. Section 6 discusses how the indexes described below may be estimated fairly using a validation technique.

5.1. General notions

In the simplest case, when the response being predicted is a continuous variable that is measured completely (as distinct from *censored* measurements caused by termination of follow-up before all subjects have had the outcome of interest), one commonly used measure of predictive accuracy is the *expected squared error* of the estimate. This quantity is defined as the expected squared difference between predicted and observed values, that is, the average squared difference between predicted and observed values if the experiment were repeated infinitely often and new estimates were made at each replication. The expected squared error can also be expressed as the square of the *bias* of the estimate plus the *variance* of the estimate. Here bias refers to the expected value of the estimate minus the quantity being estimated, such as the mean blood pressure. The expected squared error is estimated in practice by the usual mean squared error.

There are two other terms for describing the components of predictive accuracy: *calibration* and *discrimination*. Calibration refers to the extent of bias. For example, if the average predicted mortality for a group of similar patients is 0·3 and the actual proportion dying is 0·3, the predictions are well calibrated. Discrimination measures a predictor's ability to separate patients with different responses. A weather forecaster who predicts a 0·15 chance of rain every day of the year may be well calibrated in a certain locality if the average number of days with rain is 55 per year, but the forecasts are uninformative. A discriminating forecaster would be one who assigns a wide distribution of predictions and whose predicted risks for days where rain actually occurred are larger than for dry days. If a predictive model has poor discrimination, no adjustment or

‡ Often one wishes to designate a model as 'minimally acceptable' on the basis of some statistic, but in many cases it is only possible to judge a model's accuracy relative to another model. For example, a model for the probability of death after open heart surgery may yield predicted probabilities that range from 0·001 to 0·1, so the model will not have a high correlation (say 0·13) between predicted probability and observed outcome, but it may still be useful. If that model does not fully adjust for patient risk factors, it may be inadequate for adjusting for case mix when comparing mortalities among several hospitals. A more sensitive model with a correlation of, say, 0·135 may adjust away apparent differences in mortality among hospitals.

calibration can correct the model. However, if discrimination is good, the predictor can be calibrated without sacrificing the discrimination (see Section 6 for a method for calibrating predictions without needing more data). Here, calibrating predictions means modifying them, without changing their rank order, such that the predictions are perfectly calibrated. van Houwelingen and le Cessie[34] present extensive information on predictive accuracy and model validation.

5.2. Continuous uncensored outcomes

Discrimination is related to the expected squared error and to the correlation between predicted and observed responses. In the case of ordinary multiple linear regression, discrimination can be measured by the squared multiple correlation coefficient R^2, which is defined by

$$R^2 = 1 - (n - p)\text{MSE}/(n - 1)S_Y^2, \tag{1}$$

where n is the number of patients, p is the number of parameters estimated, MSE is the mean squared error of prediction $(\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2/(n - p)$, $\hat{Y}$ = predicted Y), and S_Y^2 is the sample variance of the dependent variable. When $R^2 = 1$, the model is perfectly able to separate all patient responses based on the predictor variables, and MSE = 0.

For a continuous uncensored response Y, calibration can be assessed by a scatter plot of $\hat{Y}$ (predicted Y) versus Y, optionally using a non-parametric smoother to make trends more evident.

5.3. Discrete or censored outcomes

When the outcome variable is dichotomous and predictions are stated as probabilities that an event will occur, calibration and discrimination are more informative than expected squared error alone in measuring accuracy.

One way to assess calibration of probability predictions is to form subgroups of patients and check for bias by comparing predicted and observed responses (reference 29, pp. 140–145). For example, one may group by deciles of predicted probabilities and plot the mean response (proportion with the outcome) versus the mean prediction in the decile group. However, the groupings can be quite arbitrary. Another approach is to use a smoother such as the 'super smoother'[35] or a scatterplot smoother[36] to obtain a non-parametric estimate of the relationship between $\hat{Y}$ and Y. Such smoothers work well even when Y is binary. The resulting smoothed function is a nonparametric calibration or reliability curve. Smoothers operate on the raw data $(\hat{Y}, Y)$ and do not require grouping $\hat{Y}$, but they do require one to choose a smoothing parameter or bandwidth.

As an example, consider a 7-variable binary logistic regression model to predict the probability that a certain disease is present. The model was developed on a simulated 200-subject dataset of whom 93 had a final diagnosis that is positive. While fixing the intercept and 7 regression coefficients estimated from the training sample, predictive probabilities of disease were computed for each of 200 subjects in a separate sample, of whom 104 had the disease. The non-parametric calibration curve was estimated using a local least squares scatterplot smoother[36] with the S-Plus function lowess,[37] using the 'no iteration' option. The smoothed calibration graph is shown in Figure 1. Also shown is the proportion of patients with disease, grouped by intervals of predicted probability each containing 50 patients.

Note the typical regression to the mean effect caused by overfitting: predicted probabilities in the range of 0·3 to 0·5 are too low. Actual probabilities are closer to the mean (104/200 = 0·52).