Prediction and Explanation are Different

Introduction

Depending on the goals of a project, multiple linear regression (MR) can be applied to explain or to predict. If it is the former, MR is used to examine the relationship between a dependent variable and multiple variables in a sample to ***explain how the expected value of the dependent variable relates to the predictors being studied***. If it is the latter, MR is used to generate an equation which may be used to ***predict an individual outcome in the future***. The former is retrospective in focus explaining what has been observed. The latter is prospective in focus predicting something yet to be observed. It is very likely different variables will be used in the resulting models. A hypothetical example is given below to characterize the difference between the two approaches.

Explanation

Assume the goal is to ***explain*** what affects the achievement test scores of twelfth grade students. Assume there is literature available indicating different variables are expected to be instrumental in affecting the level of achievement attained. These could be race, sex, parental education, family socioeconomic status, etc. A study is planned which includes the relevant factors theoretically considered as likely drivers of achievement as well as some considered potentially important by the investigator. Data are collected and a MR analysis is conducted ***to test hypotheses*** regarding each variable. The variables found to be statistically significant are evaluated for relative importance. The resulting report would present these variables as instrumental factors affecting achievement for the ***population of students*** studied.

Prediction

Now suppose the goal is to ***predict*** for ***each student*** what their achievement test scores will be. The intention may be to have an equation that would be used by guidance counselors to determine which students need counseling or special programs. Variables like those used in the first study might be considered. However, it is also important to consider variables that ***correlate well*** with performance on an individual level. Variables like GPA and prior test results could be considered, but if something like the number of hours worked in a week, non-academic honors, or mode of dress correlates with the outcome, such variables could be included. It may well be difficult to acquire desirable personal information for the model. An extensive list of variables would be constructed and statistical methods for eliminating and selecting variables would be used to arrive at a MR equation. Since this equation is to be used for prediction, it would need to be ***validated*** and the ***accuracy*** of its predictions evaluated.

Some Concluding Remarks

An explanatory model may have predictive capability, but there are examples where a comparable predictive model does better. On the other hand, a predictive model, built using variables that correlate well, may not offer explanatory insight into causal mechanisms at work. Consequently, it is important to know the intended use, and to recognize the resulting models may be different. It is also essential to recognize a model is not reality. No model is perfect, but some are useful for intended purposes.