Chapter 1
Introduction

1.1 Hypothesis Testing, Estimation, and Prediction

Statistics comprises among other areas study design, hypothesis testing, estimation, and prediction. This text aims at the last area, by presenting methods that enable an analyst to develop models that will make accurate predictions of responses for *future* observations. Prediction could be considered a superset of hypothesis testing and estimation, so the methods presented here will also assist the analyst in those areas. It is worth pausing to explain how this is so.

In traditional hypothesis testing one often chooses a *null hypothesis* defined as the absence of some effect. For example, in testing whether a variable such as cholesterol is a risk factor for sudden death, one might test the null hypothesis that an increase in cholesterol does not increase the risk of death. Hypothesis testing can easily be done within the context of a statistical model, but a model is not required. When one only wishes to assess whether an effect is zero, P-values may be computed using permutation or rank (nonparametric) tests while making only minimal assumptions. But there are still reasons for preferring a model-based approach over techniques that only yield P-values.

1. Permutation and rank tests do not easily give rise to estimates of *magnitudes* of effects.
2. These tests cannot be readily extended to incorporate complexities such as cluster sampling or repeated measurements within subjects.
3. Once the analyst is familiar with a model, that model may be used to carry out many different statistical tests; there is no need to learn specific formulas to handle the special cases. The two-sample t-test is a special case of the ordinary multiple regression model having as its sole X variable a dummy variable indicating group membership. The Wilcoxon-Mann-Whitney test is a special case of the proportional odds ordinal logistic

F.E. Harrell, Jr., *Regression Modeling Strategies*, Springer Series in Statistics, DOI 10.1007/978-3-319-19425-7_1

model.[664] The analysis of variance (multiple group) test and the Kruskal–Wallis test can easily be obtained from these two regression models by using more than one dummy predictor variable.

Even without complexities such as repeated measurements, problems can arise when many hypotheses are to be tested. Testing too many hypotheses is related to fitting too many predictors in a regression model. One commonly hears the statement that "the dataset was too small to allow modeling, so we just did hypothesis tests." It is unlikely that the resulting inferences would be reliable. If the sample size is insufficient for modeling it is often insufficient for tests or estimation. This is especially true when one desires to publish an estimate of the effect corresponding to the hypothesis yielding the smallest P-value. Ordinary point estimates are known to be badly biased when the quantity to be estimated was determined by "data dredging." This can be remedied by the same kind of shrinkage used in multivariable modeling (Section 9.10).

Statistical estimation is usually model-based. For example, one might use a survival regression model to estimate the relative effect of increasing cholesterol from 200 to 250 mg/dl on the hazard of death. Variables other than cholesterol may also be in the regression model, to allow estimation of the effect of increasing cholesterol, holding other risk factors constant. But accurate estimation of the cholesterol effect will depend on how cholesterol as well as each of the adjustment variables is assumed to relate to the hazard of death. If linear relationships are incorrectly assumed, estimates will be inaccurate. Accurate estimation also depends on avoiding overfitting the adjustment variables. If the dataset contains 200 subjects, 30 of whom died, and if one adjusted for 15 "confounding" variables, the estimates would be "overadjusted" for the effects of the 15 variables, as some of their apparent effects would actually result from spurious associations with the response variable (time until death). The overadjustment would reduce the cholesterol effect. The resulting unreliability of estimates equals the degree to which the overall model fails to validate on an independent sample.

It is often useful to think of effect estimates as differences between two predicted values from a model. This way, one can account for nonlinearities and interactions. For example, if cholesterol is represented nonlinearly in a logistic regression model, predicted values on the "linear combination of X's scale" are predicted log odds of an event. The increase in log odds from raising cholesterol from 200 to 250 mg/dl is the difference in predicted values, where cholesterol is set to 250 and then to 200, and all other variables are held constant. The point estimate of the 250:200 mg/dl odds ratio is the anti-log of this difference. If cholesterol is represented nonlinearly in the model, it does not matter how many terms in the model involve cholesterol as long as the overall predicted values are obtained.

Thus when one develops a reasonable multivariable predictive model, hypothesis testing and estimation of effects are byproducts of the fitted model. So predictive modeling is often desirable even when prediction is not the main goal.

1.2 Examples of Uses of Predictive Multivariable Modeling

There is an endless variety of uses for multivariable models. Predictive models have long been used in business to forecast financial performance and to model consumer purchasing and loan pay-back behavior. In ecology, regression models are used to predict the probability that a fish species will disappear from a lake. Survival models have been used to predict product life (e.g., time to burn-out of an mechanical part, time until saturation of a disposable diaper). Models are commonly used in discrimination litigation in an attempt to determine whether race or sex is used as the basis for hiring or promotion, after taking other personnel characteristics into account.

Multivariable models are used extensively in medicine, epidemiology, biostatistics, health services research, pharmaceutical research, and related fields. The author has worked primarily in these fields, so most of the examples in this text come from those areas. In medicine, two of the major areas of application are diagnosis and prognosis. There models are used to predict the probability that a certain type of patient will be shown to have a specific disease, or to predict the time course of an already diagnosed disease. In observational studies in which one desires to compare patient outcomes between two or more treatments, multivariable modeling is very important because of the biases caused by nonrandom treatment assignment. Here the simultaneous effects of several 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. A newer technique for more aggressively adjusting for nonrandom treatment assignment, the *propensity score*,[116,530] provides yet another opportunity for multivariable modeling (see Section 10.1.4). The propensity score is merely the predicted value from a multivariable model where the response variable is the exposure or the treatment actually used. The estimated propensity score is then used in a second step as an adjustment variable in the model for the response of interest.

It is not widely recognized that multivariable modeling is extremely valuable even in well-designed randomized experiments. Such studies are often designed to make *relative* comparisons of two or more treatments, using odds ratios, hazard ratios, and other measures of relative effects. But to be able to estimate *absolute* effects one must develop a multivariable model of the response variable. This model can predict, for example, the probability that a patient on treatment A with characteristics X will survive five years, or it can

predict the life expectancy for this patient. By making the same prediction for a patient on treatment B with the same characteristics, one can estimate the absolute difference in probabilities or life expectancies. This approach recognizes that low-risk patients must have less absolute benefit of treatment (lower change in outcome probability) than high-risk patients,[351] a fact that has been ignored in many clinical trials. Another reason for multivariable modeling in randomized clinical trials is that when the basic response model is nonlinear (e.g., logistic, Cox, parametric survival models), the unadjusted estimate of the treatment effect is not correct if there is moderate heterogeneity of subjects, even with perfect balance of baseline characteristics across the treatment groups.[a][9, 24, 198, 588] So even when investigators are interested in simple comparisons of two groups' responses, multivariable modeling can be advantageous and sometimes mandatory.

Cost-effectiveness analysis is becoming increasingly used in health care research, and the "effectiveness" (denominator of the cost-effectiveness ratio) is always a measure of absolute effectiveness. As absolute effectiveness varies dramatically with the risk profiles of subjects, it must be estimated for individual subjects using a multivariable model[90, 344].

1.3 Prediction vs. Classification

For problems ranging from bioinformatics to marketing, many analysts desire to develop "classifiers" instead of developing predictive models. Consider an optimum case for classifier development, in which the response variable is binary, the two levels represent a sharp dichotomy with no gray zone (e.g., complete success vs. total failure with no possibility of a partial success), the user of the classifier is forced to make one of the two choices, the cost of misclassification is the same for every future observation, and the ratio of the cost of a false positive to that of a false negative equals the (often hidden) ratio implied by the analyst's classification rule. Even if all of those conditions are met, classification is still inferior to probability modeling for driving the development of a predictive instrument or for estimation or hypothesis testing. It is far better to use the full information in the data to develop a probability model, then develop classification rules on the basis of estimated probabilities. At the least, this forces the analyst to use a proper accuracy score[219] in finding or weighting data features.

When the dependent variable is ordinal or continuous, classification through forced up-front dichotomization in an attempt to simplify the problem results in arbitrariness and major information loss even when the optimum cut point

[a] For example, unadjusted odds ratios from 2×2 tables are different from adjusted odds ratios when there is variation in subjects' risk factors within each treatment group, even when the distribution of the risk factors is identical between the two groups.

(the median) is used. Dichtomizing the outcome at a different point may require a many-fold increase in sample size to make up for the lost information[187]. In the area of medical diagnosis, it is often the case that the disease is really on a continuum, and predicting the severity of disease (rather than just its presence or absence) will greatly increase power and precision, not to mention making the result less arbitrary.

It is important to note that two-group classification represents an artificial forced choice. It is not often the case that the user of the classifier needs to be limited to two possible actions. The best option for many subjects may be to refuse to make a decision or to obtain more data (e.g., order another medical diagnostic test). A gray zone can be helpful, and predictions include gray zones automatically.

Unlike prediction (e.g., of absolute risk), classification implicitly uses utility functions (also called loss or cost functions, e.g., cost of a false positive classification). Implicit utility functions are highly problematic. First, it is well known that the utility function depends on variables that are not predictive of outcome and are not collected (e.g., subjects' preferences) that are available only at the decision point. Second, the approach assumes every subject has the same utility function[b]. Third, the analyst presumptuously assumes that the subject's utility coincides with his own.

Formal decision analysis uses subject-specific utilities and optimum predictions based on all available data[62,74,183,210,219,642c]. It follows that receiver

[b] Simple examples to the contrary are the less weight given to a false negative diagnosis of cancer in the elderly and the aversion of some subjects to surgery or chemotherapy.

[c] To make an optimal decision you need to know all relevant data about an individual (used to estimate the probability of an outcome), and the utility (cost, loss function) of making each decision. Sensitivity and specificity do not provide this information. For example, if one estimated that the probability of a disease given age, sex, and symptoms is 0.1 and the "cost"of a false positive equaled the "cost" of a false negative, one would act as if the person does not have the disease. Given other utilities, one would make different decisions. If the utilities are unknown, one gives the best estimate of the probability of the outcome to the decision maker and let her incorporate her own unspoken utilities in making an optimum decision for her.

Besides the fact that cutoffs that are not individualized do not apply to individuals, only to groups, individual decision making does not utilize sensitivity and specificity. For an individual we can compute $\text{Prob}(Y = 1|X = x)$; we don't care about $\text{Prob}(Y = 1|X > c)$, and an individual having $X = x$ would be quite puzzled if she were given $\text{Prob}(X > c|\text{future unknown Y})$ when she already knows $X = x$ so X is no longer a random variable.

Even when group decision making is needed, sensitivity and specificity can be bypassed. For mass marketing, for example, one can rank order individuals by the estimated probability of buying the product, to create a lift curve. This is then used to target the k most likely buyers where k is chosen to meet total program cost constraints.

operating characteristic curve (ROC[d]) analysis is misleading except for the special case of mass one-time group decision making with unknown utilities (e.g., launching a flu vaccination program).

An analyst's goal should be the development of the most accurate and reliable predictive model or the best model on which to base estimation or hypothesis testing. In the vast majority of cases, classification is the task of the user of the predictive model, at the point in which utilities (costs) and preferences are known.

1.4 Planning for Modeling

When undertaking the development of a model to predict a response, one of the first questions the researcher must ask is "will this model actually be used?" Many models are never used, for several reasons[522] including: (1) it was not deemed relevant to make predictions in the setting envisioned by the authors; (2) potential users of the model did not trust the relationships, weights, or variables used to make the predictions; and (3) the variables necessary to make the predictions were not routinely available.

Once the researcher convinces herself that a predictive model is worth developing, there are many study design issues to be addressed.[18,378] Models are often developed using a "convenience sample," that is, a dataset that was not collected with such predictions in mind. The resulting models are often fraught with difficulties such as the following.

1. The most important predictor or response variables may not have been collected, tempting the researchers to make do with variables that do not capture the real underlying processes.
2. The subjects appearing in the dataset are ill-defined, or they are not representative of the population for which inferences are to be drawn; similarly, the data collection sites may not represent the kind of variation in the population of sites.
3. Key variables are missing in large numbers of subjects.
4. Data are not missing at random; for example, data may not have been collected on subjects who dropped out of a study early, or on patients who were too sick to be interviewed.
5. Operational definitions of some of the key variables were never made.
6. Observer variability studies may not have been done, so that the reliability of measurements is unknown, or there are other kinds of important measurement errors.

A predictive model will be more accurate, as well as useful, when data collection is planned prospectively. That way one can design data collection

[d] The ROC curve is a plot of sensitivity vs. one minus specificity as one varies a cutoff on a continuous predictor used to make a decision.

instruments containing the necessary variables, and all terms can be given standard definitions (for both descriptive and response variables) for use at all data collection sites. Also, steps can be taken to minimize the amount of missing data.

In the context of describing and modeling health outcomes, Iezzoni[317] has an excellent discussion of the dimensions of risk that should be captured by variables included in the model. She lists these general areas that should be quantified by predictor variables:

1. age,
2. sex,
3. acute clinical stability,
4. principal diagnosis,
5. severity of principal diagnosis,
6. extent and severity of comorbidities,
7. physical functional status,
8. psychological, cognitive, and psychosocial functioning,
9. cultural, ethnic, and socioeconomic attributes and behaviors,
10. health status and quality of life, and
11. patient attitudes and preferences for outcomes.

Some baseline covariates to be sure to capture in general include

1. a baseline measurement of the response variable,
2. the subject's most recent status,
3. the subject's trajectory as of time zero or past levels of a key variable,
4. variables explaining much of the variation in the response, and
5. more subtle predictors whose distributions strongly differ between the levels of a key variable of interest in an observational study.

Many things can go wrong in statistical modeling, including the following.

1. The process generating the data is not stable.
2. The model is misspecified with regard to nonlinearities or interactions, or there are predictors missing.
3. The model is misspecified in terms of the transformation of the response variable or the model's distributional assumptions.
4. The model contains discontinuities (e.g., by categorizing continuous predictors or fitting regression shapes with sudden changes) that can be gamed by users.
5. Correlations among subjects are not specified, or the correlation structure is misspecified, resulting in inefficient parameter estimates and overconfident inference.
6. The model is overfitted, resulting in predictions that are too extreme or positive associations that are false.

7. The user of the model relies on predictions obtained by extrapolating to combinations of predictor values well outside the range of the dataset used to develop the model.
8. Accurate and discriminating predictions can lead to behavior changes that make future predictions inaccurate.

1.4.1 Emphasizing Continuous Variables

When designing the data collection it is important to emphasize the use of continuous variables over categorical ones. Some categorical variables are subjective and hard to standardize, and on the average they do not contain the same amount of statistical information as continuous variables. Above all, it is unwise to categorize naturally continuous variables during data collection,[e] as the original values can then not be recovered, and if another researcher feels that the (arbitrary) cutoff values were incorrect, other cutoffs cannot be substituted. Many researchers make the mistake of assuming that categorizing a continuous variable will result in less measurement error. This is a false assumption, for if a subject is placed in the wrong interval this will be as much as a 100% error. Thus the magnitude of the error multiplied by the probability of an error is no better with categorization.

1.5 Choice of the Model

The actual method by which an underlying statistical model should be chosen by the analyst is not well developed. A. P. Dawid is quoted in Lehmann[397] as saying the following.

> Where do probability models come from? To judge by the resounding silence over this question on the part of most statisticians, it seems highly embarrassing. In general, the theoretician is happy to accept that his abstract probability triple (Ω, A, P) was found under a gooseberry bush, while the applied statistician's model "just growed".

In biostatistics, epidemiology, economics, psychology, sociology, and many other fields it is seldom the case that subject matter knowledge exists that would allow the analyst to pre-specify a model (e.g., Weibull or log-normal survival model), a transformation for the response variable, and a structure

[e] An exception may be sensitive variables such as income level. Subjects may be more willing to check a box corresponding to a wide interval containing their income. It is unlikely that a reduction in the probability that a subject will inflate her income will offset the loss of precision due to categorization of income, but there will be a decrease in the number of refusals. This reduction in missing data can more than offset the lack of precision.

for how predictors appear in the model (e.g., transformations, addition of nonlinear terms, interaction terms). Indeed, some authors question whether the notion of a true model even exists in many cases.[100] We are for better or worse forced to develop models empirically in the majority of cases. Fortunately, careful and objective validation of the accuracy of model predictions against observable responses can lend credence to a model, if a good validation is not merely the result of overfitting (see Section 5.3).

There are a few general guidelines that can help in choosing the basic form of the statistical model.

1. The model must use the data efficiently. If, for example, one were interested in predicting the probability that a patient with a specific set of characteristics would live five years from diagnosis, an inefficient model would be a binary logistic model. A more efficient method, and one that would also allow for losses to follow-up before five years, would be a semiparametric (rank based) or parametric survival model. Such a model uses individual times of events in estimating coefficients, but it can easily be used to estimate the probability of surviving five years. As another example, if one were interested in predicting patients' quality of life on a scale of excellent, very good, good, fair, and poor, a polytomous (multinomial) categorical response model would not be efficient as it would not make use of the ordering of responses.
2. Choose a model that fits overall structures likely to be present in the data. In modeling survival time in chronic disease one might feel that the importance of most of the risk factors is constant over time. In that case, a proportional hazards model such as the Cox or Weibull model would be a good initial choice. If on the other hand one were studying acutely ill patients whose risk factors wane in importance as the patients survive longer, a model such as the log-normal or log-logistic regression model would be more appropriate.
3. Choose a model that is robust to problems in the data that are difficult to check. For example, the Cox proportional hazards model and ordinal logistic models are not affected by monotonic transformations of the response variable.
4. Choose a model whose mathematical form is appropriate for the response being modeled. This often has to do with minimizing the need for interaction terms that are included only to address a basic lack of fit. For example, many researchers have used ordinary linear regression models for binary responses, because of their simplicity. But such models allow predicted probabilities to be outside the interval $[0, 1]$, and strange interactions among the predictor variables are needed to make predictions remain in the legal range.
5. Choose a model that is readily extendible. The Cox model, by its use of stratification, easily allows a few of the predictors, especially if they are categorical, to violate the assumption of equal regression coefficients over

time (proportional hazards assumption). The continuation ratio ordinal logistic model can also be generalized easily to allow for varying coefficients of some of the predictors as one proceeds across categories of the response.

R. A. Fisher as quoted in Lehmann[397] had these suggestions about model building: "(a) We must confine ourselves to those forms which we know how to handle," and (b) "More or less elaborate forms will be suitable according to the volume of the data." Ameen [100, p. 453] stated that a good model is "(a) satisfactory in performance relative to the stated objective, (b) logically sound, (c) representative, (d) questionable and subject to on-line interrogation, (e) able to accommodate external or expert information and (f) able to convey information."

It is very typical to use the data to make decisions about the form of the model as well as about how predictors are represented in the model. Then, once a model is developed, the entire modeling process is routinely forgotten, and statistical quantities such as standard errors, confidence limits, P-values, and R^2 are computed as if the resulting model were entirely pre-specified. However, Faraway,[186] Draper,[163] Chatfield,[100] Buckland et al.[80] and others have written about the severe problems that result from treating an empirically derived model as if it were pre-specified and as if it were the correct model. As Chatfield states [100, p. 426]:"It is indeed strange that we often admit model uncertainty by searching for a best model but then ignore this uncertainty by making inferences and predictions as if certain that the best fitting model is actually true."

Stepwise variable selection is one of the most widely used and abused of all data analysis techniques. Much is said about this technique later (see Section 4.3), but there are many other elements of model development that will need to be accounted for when making statistical inferences, and unfortunately it is difficult to derive quantities such as confidence limits that are properly adjusted for uncertainties such as the data-based choice between a
4 Weibull and a log-normal regression model.

Ye[678] developed a general method for estimating the "generalized degrees of freedom" (GDF) for any "data mining" or model selection procedure based on least squares. The GDF is an extremely useful index of the amount of "data dredging" or overfitting that has been done in a modeling process. It is also useful for estimating the residual variance with less bias. In one example, Ye developed a regression tree using recursive partitioning involving 10 candidate predictor variables on 100 observations. The resulting tree had 19 nodes and GDF of 76. The usual way of estimating the residual variance involves dividing the pooled within-node sum of squares by $100 - 19$, but Ye showed that dividing by $100 - 76$ instead yielded a much less biased (and much higher) estimate of σ^2. In another example, Ye considered stepwise variable selection using 20 candidate predictors and 22 observations. When there is no true association between any of the predictors and the response, Ye found that GDF = 14.1 for a strategy that selected the best five-variable
5 model.

Given that the choice of the model has been made (e.g., a log-normal model), penalized maximum likelihood estimation has major advantages in the battle between making the model fit adequately and avoiding overfitting (Sections 9.10 and 13.4.7). Penalization lessens the need for model selection.

1.6 Further Reading

1 Briggs and Zaretzki[74] eloquently state the problem with ROC curves and the areas under them (AUC):

> Statistics such as the AUC are not especially relevant to someone who must make a decision about a particular x_c. ... ROC curves lack or obscure several quantities that are necessary for evaluating the operational effectiveness of diagnostic tests. ... ROC curves were first used to check how radio *receivers* (like radar receivers) operated over a range of frequencies. ... This is not how must ROC curves are used now, particularly in medicine. The receiver of a diagnostic measurement ... wants to make a decision based on some x_c, and is not especially interested in how well he would have done had he used some different cutoff.

In the discussion to their paper, David Hand states

> When integrating to yield the overall AUC measure, it is necessary to decide what weight to give each value in the integration. The AUC implicitly does this using a weighting derived empirically from the data. This is nonsensical. The relative importance of misclassifying a case as a noncase, compared to the reverse, cannot come from the data itself. It must come externally, from considerations of the severity one attaches to the different kinds of misclassifications.

AUC, only because it equals the concordance probability in the binary Y case, is still often useful as a predictive discrimination measure.

2 More severe problems caused by dichotomizing continuous variables are discussed in [13, 17, 45, 82, 185, 294, 379, 521, 597].

3 See the excellent editorial by Mallows[434] for more about model choice. See Breiman and discussants[67] for an interesting debate about the use of data models vs. algorithms. This material also covers interpretability vs. predictive accuracy and several other topics.

4 See [15, 80, 100, 163, 186, 415] for information about accounting for model selection in making final inferences. Faraway[186] demonstrated that the bootstrap has good potential in related although somewhat simpler settings, and Buckland et al.[80] developed a promising bootstrap weighting method for accounting for model uncertainty.

5 Tibshirani and Knight[611] developed another approach to estimating the generalized degrees of freedom. Luo et al.[430] developed a way to add noise of known variance to the response variable to tune the stopping rule used for variable selection. Zou et al.[689] showed that the lasso, an approach that simultaneously selects variables and shrinks coefficients, has a nice property. Since it uses penalization (shrinkage), an unbiased estimate of its effective number of degrees of freedom is the number of nonzero regression coefficients in the final model.