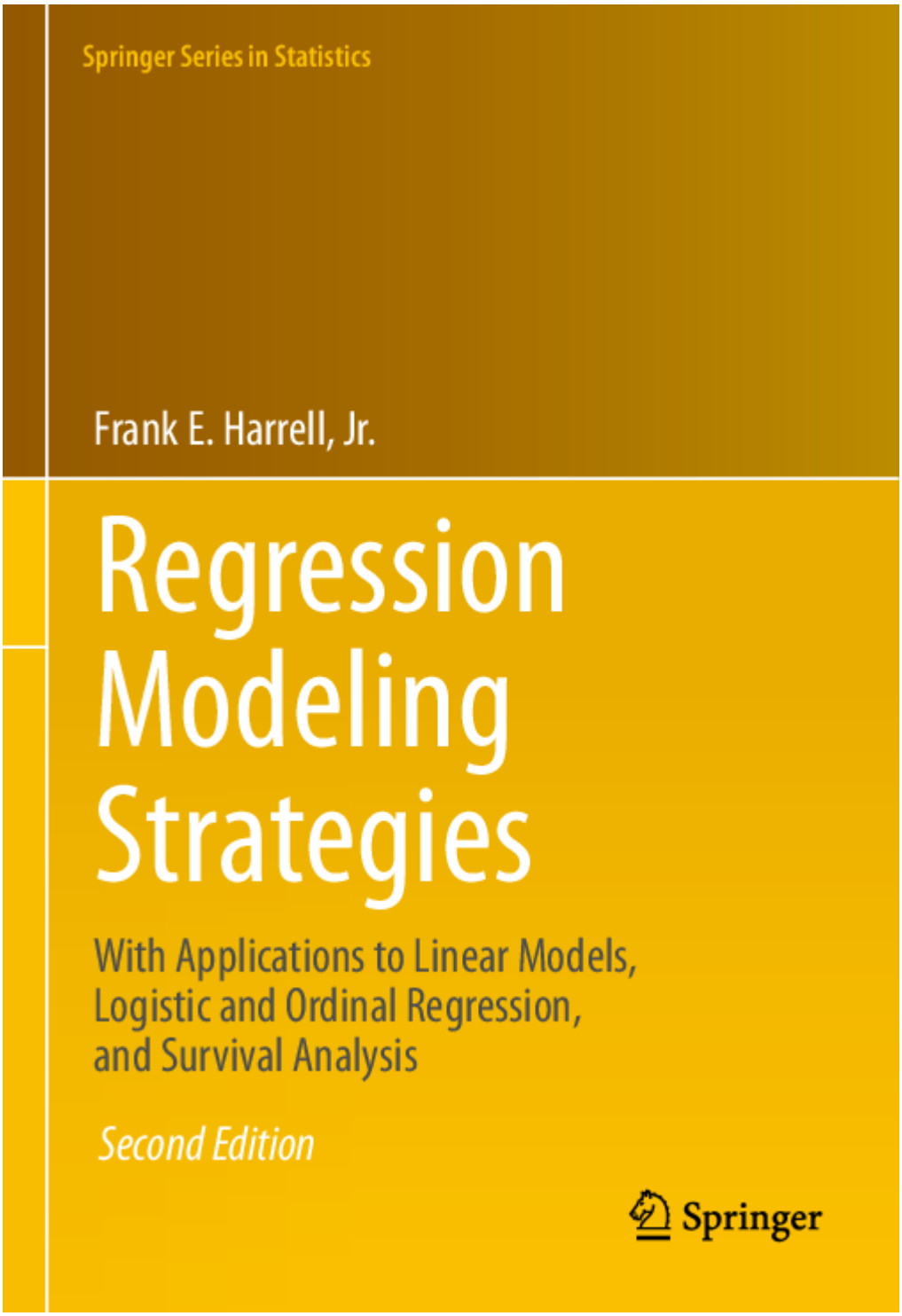

Springer Series in Statistics
Frank E. Harrell, Jr.
Regression Modeling Strategies
With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis
Second Edition
Springer

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*

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.[a9,24,198,588] 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 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

[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.

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 Weibull and a log-normal regression model. 4

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 model. 5

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.

Chapter 2
General Aspects of Fitting Regression Models

2.1 Notation for Multivariable Regression Models

The ordinary multiple linear regression model is frequently used and has parameters that are easily interpreted. In this chapter we study a general class of regression models, those stated in terms of a weighted sum of a set of independent or predictor variables. It is shown that after linearizing the model with respect to the predictor variables, the parameters in such regression models are also readily interpreted. Also, all the designs used in ordinary linear regression can be used in this general setting. These designs include analysis of variance (ANOVA) setups, interaction effects, and nonlinear effects. Besides describing and interpreting general regression models, this chapter also describes, in general terms, how the three types of assumptions of regression models can be examined.

First we introduce notation for regression models. Let Y denote the response (dependent) variable, and let $X = X_1, X_2, \ldots, X_p$ denote a list or vector of predictor variables (also called covariables or independent, descriptor, or concomitant variables). These predictor variables are assumed to be constants for a given individual or subject from the population of interest. Let $\beta = \beta_0, \beta_1, \ldots, \beta_p$ denote the list of regression coefficients (parameters). β_0 is an optional intercept parameter, and $\beta_1, \ldots, \beta_p$ are weights or regression coefficients corresponding to $X_1, \ldots, X_p$. We use matrix or vector notation to describe a weighted sum of the Xs:

$$X\beta = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p, \tag{2.1}$$

where there is an implied $X_0 = 1$.

A regression model is stated in terms of a connection between the predictors X and the response Y. Let $C(Y|X)$ denote a property of the distribution of Y given X (as a function of X). For example, $C(Y|X)$ could be $E(Y|X)$,

F.E. Harrell, Jr., *Regression Modeling Strategies*, Springer Series in Statistics, DOI 10.1007/978-3-319-19425-7_2

the expected value or average of Y given X, or $C(Y|X)$ could be the probability that $Y = 1$ given X (where $Y = 0$ or 1).

2.2 Model Formulations

We define a regression function as a function that describes interesting properties of Y that may vary across individuals in the population. X describes the list of factors determining these properties. Stated mathematically, a general regression model is given by

$$C(Y|X) = g(X). \tag{2.2}$$

We restrict our attention to models that, after a certain transformation, are linear in the unknown parameters, that is, models that involve X only through a weighted sum of all the Xs. The *general linear regression model* is given by

$$C(Y|X) = g(X\beta). \tag{2.3}$$

For example, the ordinary linear regression model is

$$C(Y|X) = E(Y|X) = X\beta, \tag{2.4}$$

and given X, Y has a normal distribution with mean $X\beta$ and constant variance σ^2. The binary logistic regression model[129,647] is

$$C(Y|X) = \text{Prob}\{Y = 1|X\} = (1 + \exp(-X\beta))^{-1}, \tag{2.5}$$

where Y can take on the values 0 and 1. In general the model, when stated in terms of the property $C(Y|X)$, may not be linear in $X\beta$; that is $C(Y|X) = g(X\beta)$, where $g(u)$ is nonlinear in u. For example, a regression model could be $E(Y|X) = (X\beta)^{.5}$. The model may be made linear in the unknown parameters by a transformation in the property $C(Y|X)$:

$$h(C(Y|X)) = X\beta, \tag{2.6}$$

where $h(u) = g^{-1}(u)$, the inverse function of g. As an example consider the binary logistic regression model given by

$$C(Y|X) = \text{Prob}\{Y = 1|X\} = (1 + \exp(-X\beta))^{-1}. \tag{2.7}$$

If $h(u) = \text{logit}(u) = \log(u/(1-u))$, the transformed model becomes

$$h(\text{Prob}(Y = 1|X)) = \log(\exp(X\beta)) = X\beta. \tag{2.8}$$

The transformation $h(C(Y|X))$ is sometimes called a *link function*. Let $h(C(Y|X))$ be denoted by $C'(Y|X)$. The general linear regression model then becomes

$$C'(Y|X) = X\beta. \tag{2.9}$$

In other words, the model states that some property C' of Y, given X, is a weighted sum of the Xs ($X\beta$). In the ordinary linear regression model, $C'(Y|X) = E(Y|X)$. In the logistic regression case, $C'(Y|X)$ is the logit of the probability that $Y = 1$, $\log \text{Prob}\{Y = 1\}/[1 - \text{Prob}\{Y = 1\}]$. This is the log of the odds that $Y = 1$ versus $Y = 0$.

It is important to note that the general linear regression model has two major components: $C'(Y|X)$ and $X\beta$. The first part has to do with a property or transformation of Y. The second, $X\beta$, is the *linear regression* or *linear predictor* part. The method of least squares can sometimes be used to fit the model if $C'(Y|X) = E(Y|X)$. Other cases must be handled using other methods such as maximum likelihood estimation or nonlinear least squares.

2.3 Interpreting Model Parameters

In the original model, $C(Y|X)$ specifies the way in which X affects a property of Y. Except in the ordinary linear regression model, it is difficult to interpret the individual parameters if the model is stated in terms of $C(Y|X)$. In the model $C'(Y|X) = X\beta = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p$, the regression parameter β_j is interpreted as the change in the property C' of Y per unit change in the descriptor variable X_j, all other descriptors remaining constant[a]:

$$\beta_j = C'(Y|X_1, X_2, \ldots, X_j + 1, \ldots, X_p) - C'(Y|X_1, X_2, \ldots, X_j, \ldots, X_p). \tag{2.10}$$

In the ordinary linear regression model, for example, β_j is the change in expected value of Y per unit change in X_j. In the logistic regression model β_j is the change in log odds that $Y = 1$ per unit change in X_j. When a non-interacting X_j is a dichotomous variable or a continuous one that is linearly related to C', X_j is represented by a single term in the model and its contribution is described fully by β_j.

In all that follows, we drop the $'$ from C' and assume that $C(Y|X)$ is the property of Y that is linearly related to the weighted sum of the Xs.

[a] Note that it is not necessary to "hold constant" all other variables to be able to interpret the effect of one predictor. It is sufficient to hold constant the weighted sum of all the variables other than X_j. And in many cases it is not physically possible to hold other variables constant while varying one, e.g., when a model contains X and X^2 (David Hoaglin, personal communication).

2.3.1 Nominal Predictors

Suppose that we wish to model the effect of two or more treatments and be able to test for differences between the treatments in some property of Y. A nominal or polytomous factor such as treatment group having k levels, in which there is no definite ordering of categories, is fully described by a series of $k-1$ binary indicator variables (sometimes called *dummy variables*). Suppose that there are four treatments, J, K, L, and M, and the treatment factor is denoted by T. The model can be written as

$$\begin{aligned} C(Y|T=J) &= \beta_0 \\ C(Y|T=K) &= \beta_0+\beta_1 \\ C(Y|T=L) &= \beta_0+\beta_2 \\ C(Y|T=M) &= \beta_0+\beta_3. \end{aligned} \tag{2.11}$$

The four treatments are thus completely specified by three regression parameters and one intercept that we are using to denote treatment J, the reference treatment. This model can be written in the previous notation as

$$C(Y|T) = X\beta = \beta_0+\beta_1X_1+\beta_2X_2+\beta_3X_3, \tag{2.12}$$

where

$$\begin{aligned} X_1 &= 1 \text{ if } T=K, 0 \text{ otherwise} \\ X_2 &= 1 \text{ if } T=L, 0 \text{ otherwise} \\ X_3 &= 1 \text{ if } T=M, 0 \text{ otherwise.} \end{aligned} \tag{2.13}$$

For treatment J ($T=J$), all three Xs are zero and $C(Y|T=J)=\beta_0$. The test for any differences in the property $C(Y)$ between treatments is $H_0: \beta_1=\beta_2=\beta_3=0$.

This model is an *analysis of variance* or *k-sample*-type model. If there are other descriptor covariables in the model, it becomes an *analysis of covariance*-type model.

2.3.2 Interactions

Suppose that a model has descriptor variables X_1 and X_2 and that the effect of the two Xs cannot be separated; that is the effect of X_1 on Y depends on the level of X_2 and vice versa. One simple way to describe this *interaction* is to add the constructed variable $X_3=X_1X_2$ to the model:

$$C(Y|X) = \beta_0+\beta_1X_1+\beta_2X_2+\beta_3X_1X_2. \tag{2.14}$$

It is now difficult to interpret β_1 and β_2 in isolation. However, we may quantify the effect of a one-unit increase in X_1 if X_2 is held constant as

Table 2.1 Parameters in a simple model with interaction

Parameter	Meaning
β_0	$C(Y\|age=0, sex=m)$
β_1	$C(Y\|age=x+1, sex=m) - C(Y\|age=x, sex=m)$
β_2	$C(Y\|age=0, sex=f) - C(Y\|age=0, sex=m)$
β_3	$C(Y\|age=x+1, sex=f) - C(Y\|age=x, sex=f) -$ $[C(Y\|age=x+1, sex=m) - C(Y\|age=x, sex=m)]$

$$\begin{aligned} C(Y|X_1+1, X_2) &- C(Y|X_1, X_2) \\ &= \beta_0+\beta_1(X_1+1)+\beta_2X_2 \\ &+ \beta_3(X_1+1)X_2 \\ &- [\beta_0+\beta_1X_1+\beta_2X_2+\beta_3X_1X_2] \\ &= \beta_1+\beta_3X_2. \end{aligned} \tag{2.15}$$

Likewise, the effect of a one-unit increase in X_2 on C if X_1 is held constant is $\beta_2+\beta_3X_1$. Interactions can be much more complex than can be modeled with a product of two terms. If X_1 is binary, the interaction may take the form of a difference in shape (and/or distribution) of X_2 versus $C(Y)$ depending on whether $X_1=0$ or $X_1=1$ (e.g., logarithm vs. square root). When both variables are continuous, the possibilities are much greater (this case is discussed later). Interactions among more than two variables can be exceedingly complex.

2.3.3 Example: Inference for a Simple Model

Suppose we postulated the model

$$C(Y|age, sex) = \beta_0+\beta_1age+\beta_2[sex=f]+\beta_3age[sex=f],$$

where $[sex=f]$ is a 0–1 indicator variable for sex = female; the reference cell is sex = male corresponding to a zero value of the indicator variable. This is a model that assumes

1. age is linearly related to $C(Y)$ for males,
2. age is linearly related to $C(Y)$ for females, and
3. whatever distribution, variance, and independence assumptions are appropriate for the model being considered.

We are thus assuming that the interaction between age and sex is simple; that is it only alters the slope of the age effect. The parameters in the model have interpretations shown in Table 2.1. β_3 is the difference in slopes (female – male).

There are many useful hypotheses that can be tested for this model. First let's consider two hypotheses that are seldom appropriate although they are routinely tested.

1. $H_0 : \beta_1 = 0$: This tests whether age is associated with Y for males.
2. $H_0 : \beta_2 = 0$: This tests whether sex is associated with Y for zero-year olds.

Now consider more useful hypotheses. For each hypothesis we should write what is being tested, translate this to tests in terms of parameters, write the alternative hypothesis, and describe what the test has maximum power to detect. The latter component of a hypothesis test needs to be emphasized, as almost every statistical test is focused on one specific pattern to detect. For example, a test of association against an alternative hypothesis that a slope is nonzero will have maximum power when the true association is linear. If the true regression model is exponential in X, a linear regression test will have some power to detect "non-flatness" but it will not be as powerful as the test from a well-specified exponential regression effect. If the true effect is U-shaped, a test of association based on a linear model will have almost no power to detect association. If one tests for association against a quadratic (parabolic) alternative, the test will have some power to detect a logarithmic shape but it will have very little power to detect a cyclical trend having multiple "humps." In a quadratic regression model, a test of linearity against a quadratic alternative hypothesis will have reasonable power to detect a quadratic nonlinear effect but very limited power to detect a multiphase cyclical trend. Therefore in the tests in Table 2.2 keep in mind that power is maximal when linearity of the age relationship holds for both sexes. In fact it may be useful to write alternative hypotheses as, for example, "H_a : age is associated with $C(Y)$, powered to detect a *linear* relationship."

Note that if there is an interaction effect, we know that there is both an age and a sex effect. However, there can also be age or sex effects when the lines are parallel. That's why the tests of total association have 2 d.f.

2.4 Relaxing Linearity Assumption for Continuous Predictors

2.4.1 Avoiding Categorization

Relationships among variables are seldom linear, except in special cases such as when one variable is compared with itself measured at a different time. It is a common belief among practitioners who do not study bias and

efficiency in depth that the presence of non-linearity should be dealt with by chopping continuous variables into intervals. Nothing could be more disastrous.[13,14,17,45,82,185,187,215,294,300,379,446,465,521,533,559,597,646]

Table 2.2 Most Useful Tests for Linear *Age* × *Sex* Model

Null or Alternative Hypothesis	Mathematical Statement
Effect of age is independent of sex or Effect of sex is independent of age or Age and sex are additive Age effects are parallel	$H_0 : \beta_3 = 0$
Age interacts with sex Age modifies effect of sex Sex modifies effect of age Sex and age are non-additive (synergistic)	$H_a : \beta_3 \neq 0$
Age is not associated with Y Age is associated with Y Age is associated with Y for either Females or males	$H_0 : \beta_1 = \beta_3 = 0$ $H_a : \beta_1 \neq 0$ or $\beta_3 \neq 0$
Sex is not associated with Y Sex is associated with Y Sex is associated with Y for some Value of age	$H_0 : \beta_2 = \beta_3 = 0$ $H_a : \beta_2 \neq 0$ or $\beta_3 \neq 0$
Neither age nor sex is associated with Y Either age or sex is associated with Y	$H_0 : \beta_1 = \beta_2 = \beta_3 = 0$ $H_a : \beta_1 \neq 0$ or $\beta_2 \neq 0$ or $\beta_3 \neq 0$

Problems caused by dichotomization include the following.

1. Estimated values will have reduced precision, and associated tests will have reduced power.
2. Categorization assumes that the relationship between the predictor and the response is flat within intervals; this assumption is far less reasonable than a linearity assumption in most cases.
3. To make a continuous predictor be more accurately modeled when categorization is used, multiple intervals are required. The needed indicator variables will spend more degrees of freedom than will fitting a smooth relationship, hence power and precision will suffer. And because of sample size limitations in the very low and very high range of the variable, the outer intervals (e.g., outer quintiles) will be wide, resulting in significant heterogeneity of subjects within those intervals, and residual confounding.
4. Categorization assumes that there is a discontinuity in response as interval boundaries are crossed. Other than the effect of time (e.g., an instant stock price drop after bad news), there are very few examples in which such discontinuities have been shown to exist.
5. Categorization only seems to yield interpretable estimates such as odds ratios. For example, suppose one computes the odds ratio for stroke for persons with a systolic blood pressure > 160 mmHg compared with persons with a blood

pressure $\leq$ 160 mmHg. The interpretation of the resulting odds ratio will depend on the exact distribution of blood pressures in the sample (the proportion of subjects > 170, > 180, etc.). On the other hand, if blood pressure is modeled as a continuous variable (e.g., using a regression spline, quadratic, or linear effect) one can estimate the ratio of odds for *exact* settings of the predictor, e.g., the odds ratio for 200 mmHg compared with 120 mmHg.

6. Categorization does not condition on full information. When, for example, the risk of stroke is being assessed for a new subject with a known blood pressure (say 162 mmHg), the subject does not report to her physician "my blood pressure exceeds 160" but rather reports 162 mmHg. The risk for this subject will be much lower than that of a subject with a blood pressure of 200 mmHg.
7. If cutpoints are determined in a way that is not blinded to the response variable, calculation of P-values and confidence intervals requires special simulation techniques; ordinary inferential methods are completely invalid. For example, if cutpoints are chosen by trial and error in a way that utilizes the response, even informally, ordinary P-values will be too small and confidence intervals will not have the claimed coverage probabilities. The correct Monte-Carlo simulations must take into account both multiplicities and uncertainty in the choice of cutpoints. For example, if a cutpoint is chosen that minimizes the P-value and the resulting P-value is 0.05, the true type I error can easily be above 0.5[300].
8. Likewise, categorization that is not blinded to the response variable results in biased effect estimates[17,559].
9. "Optimal" cutpoints do not replicate over studies. Hollander et al.[300] state that "...the optimal cutpoint approach has disadvantages. One of these is that in almost every study where this method is applied, another cutpoint will emerge. This makes comparisons across studies extremely difficult or even impossible. Altman et al. point out this problem for studies of the prognostic relevance of the S-phase fraction in breast cancer published in the literature. They identified 19 different cutpoints used in the literature; some of them were solely used because they emerged as the 'optimal' cutpoint in a specific data set. In a meta-analysis on the relationship between cathepsin-D content and disease-free survival in node-negative breast cancer patients, 12 studies were in included with 12 different cutpoints ... Interestingly, neither cathepsin-D nor the S-phase fraction are recommended to be used as prognostic markers in breast cancer in the recent update of the American Society of Clinical Oncology." Giannoni et al.[215] demonstrated that many claimed "optimal cutpoints" are just the observed median values in the sample, which happens to optimize statistical power for detecting a separation in outcomes and have nothing to do with true outcome thresholds. Disagreements in cutpoints (which are bound to happen whenever one searches for things that do not exist) cause severe interpretation problems. One study may provide an odds ratio for comparing body mass index (BMI) > 30 with BMI $\leq$ 30, another for comparing BMI > 28 with BMI $\leq$ 28. Neither of these odds ratios has a good definition and the two estimates are not comparable.
10. Cutpoints are arbitrary and manipulatable; cutpoints can be found that can result in both positive and negative associations[646].
11. If a confounder is adjusted for by categorization, there will be residual confounding that can be explained away by inclusion of the continuous form of the predictor in the model in addition to the categories.

When cutpoints are chosen using Y, categorization represents one of those few times in statistics where both type I and type II errors are elevated.

A scientific quantity is a quantity which can be defined outside of the specifics of the current experiment. The kind of high:low estimates that result from categorizing a continuous variable are not scientific quantities; their interpretation depends on the entire sample distribution of continuous measurements within the chosen intervals.

To summarize problems with categorization it is useful to examine its effective assumptions. Suppose one assumes there is a single cutpoint c for predictor X. Assumptions implicit in seeking or using this cutpoint include (1) the relationship between X and the response Y is discontinuous at $X = c$ and only $X = c$; (2) c is correctly found as the cutpoint; (3) X vs. Y is flat to the left of c; (4) X vs. Y is flat to the right of c; (5) the "optimal" cutpoint does not depend on the values of other predictors. Failure to have these assumptions satisfied will result in great error in estimating c (because it doesn't exist), low predictive accuracy, serious lack of model fit, residual confounding, and overestimation of effects of remaining variables.

A better approach that maximizes power and that only assumes a smooth relationship is to use regression splines for predictors that are not known to predict linearly. Use of flexible parametric approaches such as this allows standard inference techniques (P-values, confidence limits) to be used, as will be described below. Before introducing splines, we consider the simplest approach to allowing for nonlinearity.

2.4.2 Simple Nonlinear Terms

If a continuous predictor is represented, say, as X_1 in the model, the model is assumed to be linear in X_1. Often, however, the property of Y of interest does not behave linearly in all the predictors. The simplest way to describe a nonlinear effect of X_1 is to include a term for $X_2 = X_1^2$ in the model:

$$C(Y|X_1) = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2. \tag{2.16}$$

If the model is truly linear in X_1, β_2 will be zero. This model formulation allows one to test H_0 : model is linear in X_1 against H_a : model is quadratic (parabolic) in X_1 by testing $H_0 : \beta_2 = 0$.

Nonlinear effects will frequently not be of a parabolic nature. If a transformation of the predictor is known to induce linearity, that transformation (e.g., $\log(X)$) may be substituted for the predictor. However, often the transformation is not known. Higher powers of X_1 may be included in the model to approximate many types of relationships, but polynomials have some undesirable properties (e.g., undesirable peaks and valleys, and the fit in one region of X can be greatly affected by data in other regions[433]) and will not adequately fit many functional forms.[156] For example, polynomials do not adequately fit logarithmic functions or "threshold" effects.

b. Write an expression for the nominal P–value for testing association using this strategy.
c. Write an expression for the actual P–value or alternatively for the type–I error if using a fixed critical value for the test of association.
d. For the same two–stage strategy consider an estimate of the effect on $C(Y|X)$ of increasing X from a to b. Write a brief symbolic algorithm for deriving a true two–sided $1-\alpha$ confidence interval for the $b:a$ effect (difference in $C(Y)$) using the bootstrap.

Table 2.4 SAT data from the College Board, 1988

% Taking (X)	SAT Mean Verbal Score (Y)	% Taking (X)	SAT Mean Verbal Score (Y)
4	482	24	440
5	498	29	460
5	513	37	448
6	498	43	441
6	511	44	424
7	479	45	417
9	480	49	422
9	483	50	441
10	475	52	408
10	476	55	412
10	487	57	400
10	494	58	401
12	474	59	430
12	478	60	433
13	457	62	433
13	485	63	404
14	451	63	424
14	471	63	430
14	473	64	431
16	467	64	437
17	470	68	446
18	464	69	424
20	471	72	420
22	455	73	432
23	452	81	436

Chapter 3
Missing Data

3.1 Types of Missing Data

There are missing data in the majority of datasets one is likely to encounter. Before discussing some of the problems of analyzing data in which some variables are missing for some subjects, we define some nomenclature. 1

Missing completely at random (MCAR)

Data are missing for reasons that are unrelated to any characteristics or responses for the subject, including the value of the missing value, were it to be known. Examples include missing laboratory measurements because of a dropped test tube (if it was not dropped because of knowledge of any measurements), a study that ran out of funds before some subjects could return for follow-up visits, and a survey in which a subject omitted her response to a question for reasons unrelated to the response she would have made or to any other of her characteristics.

Missing at random (MAR)

Data are not missing at random, but the probability that a value is missing depends on values of variables that were actually measured. As an example, consider a survey in which females are less likely to provide their personal income in general (but the likelihood of responding is independent of her actual income). If we know the sex of every subject and have income levels for some of the females, unbiased sex-specific income estimates can be made. That is because the incomes we do have for some of the females are a random sample of all females' incomes. Another way of saying that a variable is MAR

F.E. Harrell, Jr., *Regression Modeling Strategies*, Springer Series in Statistics, DOI 10.1007/978-3-319-19425-7_3

is that given the values of other available variables, subjects having missing values are only randomly different from other subjects.[535] Or to paraphrase Greenland and Finkle,[242] for MAR the missingness of a covariable cannot depend on unobserved covariable values; for example whether a predictor is observed cannot depend on another predictor when the latter is missing but it can depend on the latter when it is observed. MAR and MCAR data are also called *ignorable* non-responses.

Informative missing (IM)

The tendency for a variable to be missing is a function of data that are not available, including the case when data tend to be missing if their true values are systematically higher or lower. An example is when subjects with lower income levels or very high incomes are less likely to provide their personal income in an interview. IM is also called nonignorable non-response and missing not at random (MNAR).

IM is the most difficult type of missing data to handle. In many cases, there is no fix for IM nor is there a way to use the data to test for the existence of IM. External considerations must dictate the choice of missing data models, and there are few clues for specifying a model under IM. MCAR is the easiest case to handle. Our ability to correctly analyze MAR data depends on the availability of other variables (the sex of the subject in the example above). Most of the methods available for dealing with missing data assume the data are MAR. Fortunately, even though the MAR assumption is not testable, it may hold approximately if enough variables are included in the imputation models[256].

3.2 Prelude to Modeling

No matter whether one deletes incomplete cases, carefully imputes (estimates) missing data, or uses a full maximum likelihood or Bayesian techniques to incorporate partial data, it is beneficial to characterize patterns of missingness using exploratory data analysis techniques. These techniques include binary logistic models and recursive partitioning for predicting the probability that a given variable is missing. Patterns of missingness should be reported to help readers understand the limitations of incomplete data. If you do decide to use imputation, it is also important to describe how variables are simultaneously missing. A cluster analysis of missing value status of all the variables is useful here. This can uncover cases where imputation is not as effective. For example, if the only variable moderately related to diastolic blood pressure is systolic pressure, but both pressures are missing on the same subjects, systolic pressure cannot be used to estimate diastolic blood pressure. R functions `naclus` and `naplot` in the `Hmisc` package (see p. 142) can help detect how variables are simultaneously missing. Recursive partitioning (regression tree) algorithms (see Section 2.5) are invaluable for describing which kinds of subjects are missing on a variable. Logistic regression is also an excellent tool for this purpose. A later example (p. 302) demonstrates these procedures.

It can also be helpful to explore the distribution of non-missing Y by the number of missing variables in X (including zero, i.e., complete cases on X).

3.3 Missing Values for Different Types of Response Variables

When the response variable Y is collected serially but some subjects drop out of the study before completion, there are many ways of dealing with partial information[42,412,480] including multiple imputation in phases,[381] or efficiently analyzing all available serial data using a full likelihood model. When Y is the time until an event, there are actually no missing values of Y but follow-up will be curtailed for some subjects. That leaves the case where the response is completely measured once.

It is common practice to discard subjects having missing Y. Before doing so, at minimum an analysis should be done to characterize the tendency for Y to be missing, as just described. For example, logistic regression or recursive partitioning can be used to predict whether Y is missing and to test for systematic tendencies as opposed to Y being missing completely at random. In many models, though, more efficient and less biased estimates of regression coefficients can be made by also utilizing observations missing on Y that are non-missing on X. Hence there is a definite place for imputation of Y. von Hippel[645] found advantages of using all variables to impute all others, and once imputation is finished, discarding those observations having missing Y. However if missing Y values are MCAR, up-front deletion of cases having missing Y may sometimes be preferred, as imputation requires correct specification of the imputation model.

3.4 Problems with Simple Alternatives to Imputation

Incomplete predictor information is a very common missing data problem. Statistical software packages use casewise deletion in handling missing predictors; that is, any subject having *any* predictor or Y missing will be excluded from a regression analysis. Casewise deletion results in regression coefficient estimates that can be terribly biased, imprecise, or both[353]. First consider an example where bias is the problem. Suppose that the response is death and

the predictors are age, sex, and blood pressure, and that age and sex were recorded for every subject. Suppose that blood pressure was not measured for a fraction of 0.10 of the subjects, and the most common reason for not obtaining a blood pressure was that the subject was about to die. Deletion of these very sick patients will cause a major bias (downward) in the model's intercept parameter. In general, casewise deletion will bias the estimate of the model's intercept parameter (as well as others) when the probability of a case being incomplete is related to Y and not just to X [422, Example 3.3]. van der Heijden et al.[628] discuss how complete case analysis (casewise deletion) usually assumes MCAR.

Now consider an example in which casewise deletion of incomplete records is inefficient. The inefficiency comes from the reduction of sample size, which causes standard errors to increase,[162] confidence intervals to widen, and power of tests of association and tests of lack of fit to decrease. Suppose that the response is the presence of coronary artery disease and the predictors are age, sex, LDL cholesterol, HDL cholesterol, blood pressure, triglyceride, and smoking status. Suppose that age, sex, and smoking are recorded for all subjects, but that LDL is missing in 0.18 of the subjects, HDL is missing in 0.20, and triglyceride is missing in 0.21. Assume that all missing data are MCAR and that all of the subjects missing LDL are also missing HDL and that overall 0.28 of the subjects have one or more predictors missing and hence would be excluded from the analysis. If total cholesterol were known on every subject, even though it does not appear in the model, it (along perhaps with age and sex) can be used to estimate (*impute*) LDL and HDL cholesterol and triglyceride, perhaps using regression equations from other studies. Doing the analysis on a "filled in" dataset will result in more precise estimates because the sample size would then include the other 0.28 of the subjects.

In general, observations should only be discarded if the MCAR assumption is justified, there is a rarely missing predictor of overriding importance that cannot be reliably imputed from other information, or if the fraction of observations excluded is very small and the original sample size is large. Even then, there is no advantage of such deletion other than saving analyst time. If a predictor is MAR but its missingness depends on Y, casewise deletion is biased.

The first blood pressure example points out why it can be dangerous to handle missing values by adding a dummy variable to the model. Many analysts would set missing blood pressures to a constant (it doesn't matter which constant) and add a variable to the model such as `is.na(blood.pressure)` in R notation. The coefficient for the latter dummy variable will be quite large in the earlier example, and the model will appear to have great ability to predict death. This is because some of the left-hand side of the model contaminates the right-hand side; that is, `is.na(blood.pressure)` is correlated with death. For categorical variables, another common practice is to add a new category to denote missing, adding one more degree of freedom to the

predictor and changing its meaning.[a] Jones[326], Allison [12, pp. 9–11], Donders et al.[161], Knol et al.[353] and van der Heijden et al.[628] describe why both of these missing-indicator methods are invalid even when MCAR holds.

3.5 Strategies for Developing an Imputation Model

Except in special circumstances that usually involve only very simple models, the primary alternative to deleting incomplete observations is imputation of the missing values. Many non-statisticians find the notion of estimating data distasteful, but the way to think about imputation of missing values is that "making up" data is better than discarding valuable data. It is especially distressing to have to delete subjects who are missing on an adjustment variable when a major variable of interest is not missing. So one goal of imputation is to use as much information as possible for examining any one predictor's adjusted association with Y. The overall goal of imputation is to preserve the information and meaning of the non-missing data.

At this point the analyst must make some decisions about the information to use in computing predicted values for missing values.

1. Imputation of missing values for one of the variables can ignore all other information. Missing values can be filled in by sampling non-missing values of the variable, or by using a constant such as the median or mean non-missing value.
2. Imputation algorithms can be based only on external information not otherwise used in the model for Y in addition to variables included in later modeling. For example, family income can be imputed on the basis of location of residence when such information is to remain confidential for other aspects of the analysis or when such information would require too many degrees of freedom to be spent in the ultimate response model.
3. Imputations can be derived by only analyzing interrelationships among the Xs.
4. Imputations can use relationships among the Xs and between X and Y.
5. Imputations can use X, Y, and auxiliary variables not in the model predicting Y.
6. Imputations can take into account the reason for non-response if known.

The model to estimate the missing values in a sometimes-missing (target) variable should include all variables that are either

[a] This may work if values are "missing" because of "not applicable", e.g. one has a measure of marital happiness, dichotomized as high or low, but the sample contains some unmarried people. One could have a 3-category variable with values high, low, and unmarried (Paul Allison, IMPUTE e-mail list, 4Jul09).

Chapter 4
Multivariable Modeling Strategies

Chapter 2 dealt with aspects of modeling such as transformations of predictors, relaxing linearity assumptions, modeling interactions, and examining lack of fit. Chapter 3 dealt with missing data, focusing on utilization of incomplete predictor information. All of these areas are important in the overall scheme of model development, and they cannot be separated from what is to follow. In this chapter we concern ourselves with issues related to the whole model, with emphasis on deciding on the amount of complexity to allow in the model and on dealing with large numbers of predictors. The chapter concludes with three default modeling strategies depending on whether the goal is prediction, estimation, or hypothesis testing. 1

There are many choices to be made when deciding upon a global modeling strategy, including choice between

- parametric and nonparametric procedures
- parsimony and complexity
- parsimony and good discrimination ability
- interpretable models and black boxes.

This chapter addresses some of these issues. One general theme of what follows is the idea that in statistical inference when a method is capable of worsening performance of an estimator or inferential quantity (i.e., when the method is not systematically biased in one's favor), the analyst is allowed to benefit from the method. Variable selection is an example where the analysis is systematically tilted in one's favor by directly selecting variables on the basis of P-values of interest, and all elements of the final result (including regression coefficients and P-values) are biased. On the other hand, the next section is an example of the "capitalize on the benefit when it works, and the method may hurt" approach because one may reduce the complexity of an apparently weak predictor by removing its most important component—

F.E. Harrell, Jr., *Regression Modeling Strategies*, Springer Series in Statistics, DOI 10.1007/978-3-319-19425-7_4

nonlinear effects—from how the predictor is expressed in the model. The method hides tests of nonlinearity that would systematically bias the final result.

The book's web site contains a number of simulation studies and references to others that support the advocated approaches.

4.1 Prespecification of Predictor Complexity Without Later Simplification

There are rare occasions in which one actually expects a relationship to be linear. For example, one might predict mean arterial blood pressure at two months after beginning drug administration using as baseline variables the pretreatment mean blood pressure and other variables. In this case one expects the pretreatment blood pressure to linearly relate to follow-up blood pressure, and modeling is simple[a]. In the vast majority of studies, however, there is every reason to suppose that all relationships involving nonbinary predictors are nonlinear. In these cases, the only reason to represent predictors linearly in the model is that there is insufficient information in the sample to allow us to reliably fit nonlinear relationships.[b]

Supposing that nonlinearities are entertained, analysts often use scatter diagrams or descriptive statistics to decide how to represent variables in a model. The result will often be an adequately fitting model, but confidence limits will be too narrow, P-values too small, R^2 too large, and calibration too good to be true. The reason is that the "phantom d.f." that represented potential complexities in the model that were dismissed during the subjective assessments are forgotten in computing standard errors, P-values, and R^2_{adj}. The same problem is created when one entertains several transformations (log, $\sqrt{}$, etc.) and uses the data to see which one fits best, or when one tries to simplify a spline fit to a simple transformation.

An approach that solves this problem is to prespecify the complexity with which each predictor is represented in the model, without later simplification of the model. The amount of complexity (e.g., number of knots in spline functions or order of ordinary polynomials) one can afford to fit is roughly related to the "effective sample size." It is also very reasonable to allow for greater complexity for predictors that are thought to be more powerfully related to Y. For example, errors in estimating the curvature of a regression function are consequential in predicting Y only when the regression is somewhere steep. Once the analyst decides to include a predictor in every model, it is fair to use general measures of association to quantify the predictive potential for a variable. For example, if a predictor has a low rank correlation with the response, it will not "pay" to devote many degrees of freedom to that predictor in a spline function having many knots. On the other hand, a potent predictor (with a high rank correlation) not known to act linearly might be assigned five knots if the sample size allows.

When the effective sample size available is sufficiently large so that a saturated main effects model may be fitted, a good approach to gauging predictive potential is the following.

- Let all continuous predictors be represented as restricted cubic splines with k knots, where k is the maximum number of knots the analyst entertains for the current problem.
- Let all categorical predictors retain their original categories except for pooling of very low prevalence categories (e.g., ones containing < 6 observations).
- Fit this general main effects model.
- Compute the partial χ^2 statistic for testing the association of each predictor with the response, adjusted for all other predictors. In the case of ordinary regression, convert partial F statistics to χ^2 statistics or partial R^2 values.
- Make corrections for chance associations to "level the playing field" for predictors having greatly varying d.f., e.g., subtract the d.f. from the partial χ^2 (the expected value of χ^2_p is p under H_0).
- Make certain that tests of nonlinearity are not revealed as this would bias the analyst.
- Sort the partial association statistics in descending order.

Commands in the `rms` package can be used to plot only what is needed. Here is an example for a logistic model.

```
f ← lrm(y ~ sex + race + rcs(age,5) + rcs(weight,5) +
        rcs(height,5) + rcs(blood.pressure,5))
plot(anova(f))
```

This approach, and the rank correlation approach about to be discussed, do not require the analyst to really prespecify predictor complexity, so how are they not biased in our favor? There are two reasons: the analyst has already agreed to retain the variable in the model even if the strength of the association is very low, and the assessment of association does not reveal the degree of nonlinearity of the predictor to allow the analyst to "tweak" the number of knots or to discard nonlinear terms. Any predictive ability a variable might have may be concentrated in its nonlinear effects, so using the total association measure for a predictor to save degrees of freedom by restricting the variable to be linear may result in no predictive ability. Likewise, a low association measure between a categorical variable and Y might lead the analyst to collapse some of the categories based on their frequencies. This often helps, but sometimes the categories that are so combined are the

[a] Even then, the two blood pressures may need to be transformed to meet distributional assumptions.

[b] Shrinkage (penalized estimation) is a general solution (see Section 4.5). One can always use complex models that are "penalized towards simplicity," with the amount of penalization being greater for smaller sample sizes.

which insignificant variables from the first step are not reanalyzed in later steps. Univariable screening is thus even worse than stepwise modeling as it can miss important variables that are only important after adjusting for other variables.[598] Overall, neither univariable screening nor stepwise variable selection in any way solves the problem of "too many variables, too few subjects," and they cause severe biases in the resulting multivariable model fits while losing valuable predictive information from deleting marginally significant variables.

10

The online course notes contain a simple simulation study of stepwise selection using R.

4.4 Sample Size, Overfitting, and Limits on Number of Predictors

When a model is fitted that is too complex, that it, has too many free parameters to estimate for the amount of information in the data, the worth of the model (e.g., R^2) will be exaggerated and future observed values will not agree with predicted values. In this situation, *overfitting* is said to be present, and some of the findings of the analysis come from fitting noise and not just signal, or finding spurious associations between X and Y. In this section general guidelines for preventing overfitting are given. Here we concern ourselves with the *reliability* or *calibration* of a model, meaning the ability of the model to predict future observations as well as it appeared to predict the responses at hand. For now we avoid judging whether the model is adequate for the task, but restrict our attention to the likelihood that the model has significantly overfitted the data.

11

In typical low signal–to–noise ratio situations[g], model validations on independent datasets have found the minimum training sample size for which the fitted model has an independently validated predictive discrimination that equals the apparent discrimination seen with in training sample. Similar validation experiments have considered the margin of error in estimating an absolute quantity such as event probability. Studies such as[268,270,577] have shown that in many situations a fitted regression model is likely to be reliable when the number of predictors (or *candidate* predictors if using variable selection) p is less than $m/10$ or $m/20$, where m is the "limiting sample size" given in Table 4.1. A good average requirement is $p < \frac{m}{15}$. For example, Smith et al.[577] found in one series of simulations that the expected error in Cox model predicted five–year survival probabilities was below 0.05 when $p < m/20$ for "average" subjects and below 0.10 when $p < m/20$ for "sick"

12

[g] These are situations where the true R^2 is low, unlike tightly controlled experiments and mechanistic models where signal:noise ratios can be quite high. In those situations, many parameters can be estimated from small samples, and the $\frac{m}{15}$ rule of thumb can be significantly relaxed.

Table 4.1 Limiting Sample Sizes for Various Response Variables

Type of Response Variable	Limiting Sample Size m
Continuous	n (total sample size)
Binary	$\min(n_1, n_2)$ [h]
Ordinal (k categories)	$n - \frac{1}{n^2}\sum_{i=1}^{k} n_i^3$ [i]
Failure (survival) time	number of failures [j]

subjects, where m is the number of deaths. For "average" subjects, $m/10$ was adequate for preventing expected errors > 0.1. **Note**: The number of non-intercept parameters in the model (p) is usually greater than the number of predictors. Narrowly distributed predictor variables (e.g., if all subjects' ages are between 30 and 45 or only 5% of subjects are female) will require even higher sample sizes. Note that the number of candidate variables must include all variables screened for association with the response, including nonlinear terms and interactions. Instead of relying on the rules of thumb in the table, the shrinkage factor estimate presented in the next section can be used to guide the analyst in determining how many d.f. to model (see p. 87).

Rules of thumb such as the 15:1 rule do not consider that a certain minimum sample size is needed just to estimate basic parameters such as an intercept or residual variance. This is dealt with in upcoming topics about specific models. For the case of ordinary linear regression, estimation of the residual variance is central. All standard errors, P-values, confidence intervals, and R^2 depend on having a precise estimate of σ^2. The one-sample problem of estimating a mean, which is equivalent to a linear model containing only an intercept, is the easiest case when estimating σ^2. When a sample of size n is drawn from a normal distribution, a $1-\alpha$ two-sided confidence interval for the unknown population variance σ^2 is given by

$$\frac{n-1}{\chi^2_{1-\alpha/2,n-1}} s^2 < \sigma^2 < \frac{n-1}{\chi^2_{\alpha/2,n-1}} s^2, \tag{4.1}$$

[h] See [487]. If one considers the power of a two-sample binomial test compared with a Wilcoxon test if the response could be made continuous and the proportional odds assumption holds, the effective sample size for a binary response is $3n_1n_2/n \approx 3\min(n_1, n_2)$ if n_1/n is near 0 or 1 [664, Eq. 10, 15]. Here n_1 and n_2 are the marginal frequencies of the two response levels.

[i] Based on the power of a proportional odds model two-sample test when the marginal cell sizes for the response are $n_1, \ldots, n_k$, compared with all cell sizes equal to unity (response is continuous) [664, Eq, 3]. If all cell sizes are equal, the relative efficiency of having k response categories compared with a continuous response is $1-1/k^2$ [664, Eq. 14]; for example, a five-level response is almost as efficient as a continuous one if proportional odds holds across category cutoffs.

[j] This is approximate, as the effective sample size may sometimes be boosted somewhat by censored observations, especially for non-proportional hazards methods such as Wilcoxon-type tests.[49]

procedure. A method such as cross-validation or optimization of a modified AIC must be used to choose an optimal penalty factor. An advantage of penalized estimation is that one can differentially penalize the more complex components of the model such as nonlinear or interaction effects. A drawback of ridge regression and penalized maximum likelihood is that the final model is difficult to validate unbiasedly since the optimal amount of shrinkage is usually determined by examining the entire dataset. Penalization is one of the best ways to approach the "too many variables, too little data" problem. See Section 9.10 for details.

4.6 Collinearity

When at least one of the predictors can be predicted well from the other predictors, the standard errors of the regression coefficient estimates can be inflated and corresponding tests have reduced power.[217] In stepwise variable selection, collinearity can cause predictors to compete and make the selection of "important" variables arbitrary. Collinearity makes it difficult to estimate and interpret a particular regression coefficient because the data have little information about the effect of changing one variable while holding another (highly correlated) variable constant [101, Chap. 9]. However, collinearity does not affect the joint influence of highly correlated variables when tested simultaneously. Therefore, once groups of highly correlated predictors are identified, the problem can be rectified by testing the contribution of an entire set with a multiple d.f. test rather than attempting to interpret the coefficient or one d.f. test for a single predictor.

Collinearity does not affect predictions made on the same dataset used to estimate the model parameters or on new data that have the same degree of collinearity as the original data [470, pp. 379–381] as long as extreme extrapolation is not attempted. Consider as two predictors the total and LDL cholesterols that are highly correlated. If predictions are made at the same combinations of total and LDL cholesterol that occurred in the training data, no problem will arise. However, if one makes a prediction at an inconsistent combination of these two variables, the predictions may be inaccurate and have high standard errors.

When the ordinary truncated power basis is used to derive component variables for fitting linear and cubic splines, as was described earlier, the component variables can be very collinear. It is very unlikely that this will result in any problems, however, as the component variables are connected algebraically. Thus it is not possible for a combination of, for example, x and $\max(x - 10, 0)$ to be inconsistent with each other. Collinearity problems are then more likely to result from partially redundant subsets of predictors as in the cholesterol example above.

One way to quantify collinearity is with *variance inflation factors* or *VIF*, which in ordinary least squares are diagonals of the inverse of the $X'X$ matrix scaled to have unit variance (except that a column of 1s is retained corresponding to the intercept). Note that some authors compute VIF from the correlation matrix form of the design matrix, omitting the intercept. VIF_i is $1/(1 - R_i^2)$ where R_i^2 is the squared multiple correlation coefficient between column i and the remaining columns of the design matrix. For models that are fitted with maximum likelihood estimation, the information matrix is scaled to correlation form, and VIF is the diagonal of the inverse of this scaled matrix.[147,654] Then the VIF are similar to those from a weighted correlation matrix of the original columns in the design matrix. Note that indexes such as VIF are not very informative as some variables are algebraically connected to each other. 16

The SAS `VARCLUS` procedure[539] and R `varclus` function can identify collinear predictors. Summarizing collinear variables using a summary score is more powerful and stable than arbitrary selection of one variable in a group of collinear variables (see the next section). 17

4.7 Data Reduction

The sample size need not be as large as shown in Table 4.1 if the model is to be validated independently and if you don't care that the model may fail to validate. However, it is likely that the model will be overfitted and will not validate if the sample size does not meet the guidelines. Use of data reduction methods before model development is strongly recommended if the conditions in Table 4.1 are not satisfied, and if shrinkage is not incorporated into parameter estimation. Methods such as shrinkage and data reduction reduce the effective d.f. of the model, making it more likely for the model to validate on future data. Data reduction is aimed at reducing the number of parameters to estimate in the model, without distorting statistical inference for the parameters. This is accomplished by ignoring Y during data reduction. Manipulations of X in unsupervised learning may result in a loss of information for predicting Y, but when the information loss is small, the gain in power and reduction of overfitting more than offset the loss.

Some available data reduction methods are given below.

1. Use the literature to eliminate unimportant variables.
2. Eliminate variables whose distributions are too narrow.
3. Eliminate candidate predictors that are missing in a large number of subjects, especially if those same predictors are likely to be missing for future applications of the model.
4. Use a statistical data reduction method such as incomplete principal component regression, nonlinear generalizations of principal components such

- using shrinkage (penalized estimation) to fit a large model without worrying about the sample size.

Data reduction approaches covered in the last section can yield very interpretable, stable models, but there are many decisions to be made when using a two-stage (reduction/model fitting) approach. Newer single stage approaches are evolving. These new approaches, listed on the text's web site, handle continuous predictors well, unlike recursive partitioning.

When data reduction is not required, generalized additive models[277,674] should also be considered.

4.9 Overly Influential Observations

Every observation should influence the fit of a regression model. It can be disheartening, however, if a significant treatment effect or the shape of a regression effect rests on one or two observations. Overly influential observations also lead to increased variance of predicted values, especially when variances are estimated by bootstrapping after taking variable selection into account. In some cases, overly influential observations can cause one to abandon a model, "change" the data, or get more data. Observations can be *overly influential* for several major reasons.

1. The most common reason is having too few observations for the complexity of the model being fitted. Remedies for this have been discussed in Sections 4.7 and 4.3.
2. Data transcription or data entry errors can ruin a model fit.
3. Extreme values of the predictor variables can have a great impact, even when these values are validated for accuracy. Sometimes the analyst may deem a subject so atypical of other subjects in the study that deletion of the case is warranted. On other occasions, it is beneficial to truncate measurements where the data density ends. In one dataset of 4000 patients and 2000 deaths, white blood count (WBC) ranged from 500 to 100,000 with .05 and .95 quantiles of 2755 and 26,700, respectively. Predictions from a linear spline function of WBC were sensitive to WBC > 60,000, for which there were 16 patients. There were 46 patients with WBC > 40,000. Predictions were found to be more stable when WBC was truncated at 40,000, that is, setting WBC to 40,000 if WBC > 40,000.
4. Observations containing disagreements between the predictors and the response can influence the fit. Such disagreements should not lead to discarding the observations unless the predictor or response values are erroneous as in Reason 3, or the analysis is made conditional on observations being unlike the influential ones. In one example a single extreme predictor value in a sample of size 8000 that was not on a straight line relationship with the other (X, Y) pairs caused a χ^2 of 36 for testing nonlinearity of the predictor. Remember that an imperfectly fitting model is a fact of life, and discarding the observations can inflate the model's predictive accuracy. On rare occasions, such lack of fit may lead the analyst to make changes in the model's structure, but ordinarily this is best done from the "ground up" using formal tests of lack of fit (e.g., a test of linearity or interaction).

Influential observations of the second and third kinds can often be detected by careful quality control of the data. Statistical measures can also be helpful. The most common measures that apply to a variety of regression models are *leverage*, DFBETAS, DFFIT, and DFFITS.

Leverage measures the capacity of an observation to be influential due to having extreme predictor values. Such an observation is not *necessarily* influential. To compute leverage in ordinary least squares, we define the *hat matrix* H given by

$$H = X(X'X)^{-1}X'. \tag{4.6}$$

H is the matrix that when multiplied by the response vector gives the predicted values, so it measures how an observation estimates its own predicted response. The diagonals h_{ii} of H are the leverage measures and they are not influenced by Y. It has been suggested[47] that $h_{ii} > 2(p+1)/n$ signal a high leverage point, where p is the number of columns in the design matrix X aside from the intercept and n is the number of observations. Some believe that the distribution of h_{ii} should be examined for values that are higher than typical.

DFBETAS is the change in the vector of regression coefficient estimates upon deletion of each observation in turn, scaled by their standard errors.[47] Since DFBETAS encompasses an effect for each predictor's coefficient, DFBETAS allows the analyst to isolate the problem better than some of the other measures. DFFIT is the change in the predicted $X\beta$ when the observation is dropped, and DFFITS is DFFIT standardized by the standard error of the estimate of $X\beta$. In both cases, the standard error used for normalization is recomputed each time an observation is omitted. Some classify an observation as overly influential when $|\text{DFFITS}| > 2\sqrt{(p+1)/(n-p-1)}$, while others prefer to examine the entire distribution of DFFITS to identify "outliers".[47]

Section 10.7 discusses influence measures for the logistic model, which requires maximum likelihood estimation. These measures require the use of special residuals and information matrices (in place of $X'X$).

If truly influential observations are identified using these indexes, careful thought is needed to decide how (or whether) to deal with them. Most important, there is no substitute for careful examination of the dataset before doing any analyses.[99] Spence and Garrison [581, p. 16] feel that

> Although the identification of aberrations receives considerable attention in most modern statistical courses, the emphasis sometimes seems to be on disposing of embarrassing data by searching for sources of technical error or

minimizing the influence of inconvenient data by the application of resistant methods. Working scientists often find the most interesting aspect of the analysis inheres in the lack of fit rather than the fit itself.

4.10 Comparing Two Models

Frequently one wants to choose between two competing models on the basis of a common set of observations. The methods that follow assume that the performance of the models is evaluated on a sample not used to develop either one. In this case, predicted values from the model can usually be considered as a single new variable for comparison with responses in the new dataset. These methods listed below will also work if the models are compared using the same set of data used to fit each one, as long as both models have the same effective number of (candidate or actual) parameters. This requirement prevents us from rewarding a model just because it overfits the training sample (see Section 9.8.1 for a method comparing two models of differing complexity). The methods can also be enhanced using bootstrapping or cross-validation on a single sample to get a fair comparison when the playing field is not level, for example, when one model had more opportunity for fitting or overfitting the responses.

Some of the criteria for choosing one model over the other are

1. calibration (e.g., one model is well-calibrated and the other is not),
2. discrimination,
3. face validity,
4. measurement errors in required predictors,
5. use of continuous predictors (which are usually better defined than categorical ones),
6. omission of "insignificant" variables that nonetheless make sense as risk factors,
7. simplicity (although this is less important with the availability of computers), and
8. lack of fit for specific types of subjects.

Items 3 through 7 require subjective judgment, so we focus on the other aspects. If the purpose of the models is only to rank-order subjects, calibration is not an issue. Otherwise, a model having poor calibration can be dismissed outright. Given that the two models have similar calibration, discrimination should be examined critically. Various statistical indexes can quantify discrimination ability (e.g., R^2, model χ^2, Somers' D_{xy}, Spearman's ρ, area under ROC curve—see Section 10.8). Rank measures (D_{xy}, ρ, ROC area) only measure how well predicted values can rank-order responses. For example, predicted probabilities of 0.01 and 0.99 for a pair of subjects are no better than probabilities of 0.2 and 0.8 using rank measures, if the first subject had a lower response value than the second. Therefore, rank measures such as ROC area (c index), although fine for describing a given model, may not be very sensitive in choosing between two models[118,488,493]. This is especially true when the models are strong, as it is easier to move a rank correlation from 0.6 to 0.7 than it is to move it from 0.9 to 1.0. Measures such as R^2 and the model χ^2 statistic (calculated from the predicted and observed responses) are more sensitive. Still, one may not know how to interpret the added utility of a model that boosts the R^2 from 0.80 to 0.81.

Again given that both models are equally well calibrated, discrimination can be studied more simply by examining the distribution of predicted values $\hat{Y}$. Suppose that the predicted value is the probability that a subject dies. Then high-resolution histograms of the predicted risk distributions for the two models can be very revealing. If one model assigns 0.02 of the sample to a risk of dying above 0.9 while the other model assigns 0.08 of the sample to the high risk group, the second model is more discriminating. The worth of a model can be judged by how far it goes out on a limb while still maintaining good calibration.

Frequently, one model will have a similar discrimination index to another model, but the likelihood ratio χ^2 statistic is meaningfully greater for one. Assuming corrections have been made for complexity, the model with the higher χ^2 usually has a better fit for *some* subjects, although not necessarily for the *average* subject. A crude plot of predictions from the first model against predictions from the second, possibly stratified by Y, can help describe the differences in the models. More specific analyses will determine the characteristics of subjects where the differences are greatest. Large differences may be caused by an omitted, underweighted, or improperly transformed predictor, among other reasons. In one example, two models for predicting hospital mortality in critically ill patients had the same discrimination index (to two decimal places). For the relatively small subset of patients with extremely low white blood counts or serum albumin, the model that treated these factors as continuous variables provided predictions that were very much different from a model that did not.

When comparing predictions for two models that may not be calibrated (from overfitting, e.g.), the two sets of predictions may be shrunk so as to not give credit for overfitting (see Equation 4.3).

Sometimes one wishes to compare two models that used the response variable differently, a much more difficult problem. For example, an investigator may want to choose between a survival model that used time as a continuous variable, and a binary logistic model for dead/alive at six months. Here, other considerations are also important (see Section 17.1). A model that predicts dead/alive at six months does not use the response variable effectively, and it provides no information on the chance of dying within three months.

When one or both of the models is fitted using least squares, it is useful to compare them using an error measure that was not used as the optimization criterion, such as mean absolute error or median absolute error. Mean

[26] The Hmisc abs.error.pred function computes a variety of accuracy measures based on absolute errors.

[27] Shen et al.[567] developed an "optimal approximation" method to make correct inferences after model selection.

4.14 Problems

Analyze the SUPPORT dataset (getHdata(support)) as directed below to relate selected variables to total cost of the hospitalization. Make sure this response variable is utilized in a way that approximately satisfies the assumptions of normality-based multiple regression so that statistical inferences will be accurate. See problems at the end of Chapters 3 and 7 of the text for more information. Consider as predictors mean arterial blood pressure, heart rate, age, disease group, and coma score.

1. Do an analysis to understand interrelationships among predictors, and find optimal scaling (transformations) that make the predictors better relate to each other (e.g., optimize the variation explained by the first principal component).
2. Do a redundancy analysis of the predictors, using both a less stringent and a more stringent approach to assessing the redundancy of the multiple-level variable disease group.
3. Do an analysis that helps one determine how many d.f. to devote to each predictor.
4. Fit a model, assuming the above predictors act additively, but do not assume linearity for the age and blood pressure effects. Use the truncated power basis for fitting restricted cubic spline functions with 5 knots. Estimate the shrinkage coefficient $\hat{\gamma}$.
5. Make appropriate graphical diagnostics for this model.
6. Test linearity in age, linearity in blood pressure, and linearity in heart rate, and also do a joint test of linearity simultaneously in all three predictors.
7. Expand the model to not assume additivity of age and blood pressure. Use a tensor natural spline or an appropriate restricted tensor spline. If you run into any numerical difficulties, use 4 knots instead of 5. Plot in an interpretable fashion the estimated 3-D relationship between age, blood pressure, and cost for a fixed disease group.
8. Test for additivity of age and blood pressure. Make a joint test for the overall absence of complexity in the model (linearity and additivity simultaneously).

Chapter 5
Describing, Resampling, Validating, and Simplifying the Model

5.1 Describing the Fitted Model

5.1.1 Interpreting Effects

Before addressing issues related to describing and interpreting the model and its coefficients, one can never apply too much caution in attempting to interpret results in a causal manner. Regression models are excellent tools for estimating and inferring *associations* between an X and Y given that the "right" variables are in the model. Any ability of a model to provide *causal* inference rests entirely on the faith of the analyst in the experimental design, completeness of the set of variables that are thought to measure confounding and are used for adjustment when the experiment is not randomized, lack of important measurement error, and lastly the goodness of fit of the model.

The first line of attack in interpreting the results of a multivariable analysis is to interpret the model's parameter estimates. For simple linear, additive models, regression coefficients may be readily interpreted. If there are interactions or nonlinear terms in the model, however, simple interpretations are usually impossible. Many programs ignore this problem, routinely printing such meaningless quantities as the effect of increasing age^2 by one day while holding age constant. A meaningful age change needs to be chosen, and connections between mathematically related variables must be taken into account. These problems can be solved by relying on predicted values and differences between predicted values.

Even when the model contains no nonlinear effects, it is difficult to compare regression coefficients across predictors having varying scales. Some analysts like to gauge the relative contributions of different predictors on a common scale by multiplying regression coefficients by the standard deviations of the predictors that pertain to them. This does not make sense for nonnormally distributed predictors (and regression models should not need

F.E. Harrell, Jr., *Regression Modeling Strategies*, Springer Series in Statistics, DOI 10.1007/978-3-319-19425-7_5

to make assumptions about the distributions of predictors). When a predictor is binary (e.g., sex), the standard deviation makes no sense as a scaling factor as the scale would depend on the prevalence of the predictor.[a]

It is more sensible to estimate the change in Y when X_j is changed by an amount that is subject-matter relevant. For binary predictors this is a change from 0 to 1. For many continuous predictors the interquartile range is a reasonable default choice. If the 0.25 and 0.75 quantiles of X_j are g and h, linearity holds, and the estimated coefficient of X_j is b; $b \times (h - g)$ is the effect of increasing X_j by $h - g$ units, which is a span that contains half of the sample values of X_j.

For the more general case of continuous predictors that are monotonically but not linearly related to Y, a useful point summary is the change in $X\beta$ when the variable changes from its 0.25 quantile to its 0.75 quantile. For models for which $\exp(X\beta)$ is meaningful, antilogging the predicted change in $X\beta$ results in quantities such as interquartile-range odds and hazards ratios. When the variable is involved in interactions, these ratios are estimated separately for various levels of the interacting factors. For categorical predictors, ordinary effects are computed by comparing each level of the predictor with a reference level. See Section 10.10 and Chapter 11 for tabular and graphical examples of this approach.

The model can be described using *partial effect plots* by plotting each X against $X\hat{\beta}$ holding other predictors constant. Modified versions of such plots, by nonlinearly rank-transforming the predictor axis, can show the relative importance of a predictor[336].

For an X that interacts with other factors, separate curves are drawn on the same graph, one for each level of the interacting factor.

Nomograms[40,254,339,427] provide excellent graphical depictions of all the variables in the model, in addition to enabling the user to obtain predicted values manually. Nomograms are especially good at helping the user envision interactions. See Section 10.10 and Chapter 11 for examples.

5.1.2 Indexes of Model Performance

5.1.2.1 Error Measures

Care must be taken in the choice of accuracy scores to be used in validation. Indexes can be broken down into three main areas.

Central tendency of prediction errors: These measures include mean absolute differences, mean squared differences, and logarithmic scores. An absolute measure is mean $|Y - \hat{Y}|$. The mean squared error is a commonly used and sensitive measure if there are no outliers. For the special case where Y is binary, such a measure is the Brier score, which is a quadratic proper scoring rule that combines calibration and discrimination[b]. The logarithmic proper scoring rules (related to average log-likelihood) is even more sensitive but can be harder to interpret, and can be destroyed by a single predicted probability of 0 or 1 that was incorrect.

Discrimination measures: A measure of pure discrimination is a rank correlation of $\hat{Y}$ and Y, including Spearman's ρ, Kendall's τ, and Somers' D_{xy}. When Y is binary, $D_{xy} = 2 \times (c - \frac{1}{2})$ where c is the concordance probability or area under the receiver operating characteristic curve, a linear translation of the Wilcoxon-Mann-Whitney statistic. R^2 is *mostly* a measure of discrimination, and R^2_{adj} is is a good overfitting-corrected measure, if the model is pre-specified. See Section 10.8 for more information about rank-based measures.

Discrimination measures based on variation in $\hat{Y}$: These include the regression sum of squares and the g–Index (see below).

Calibration measures: These assess absolute prediction accuracy. *Calibration–in–the–large* compares the average $\hat{Y}$ with the average Y. A *high–resolution calibration curve* or *calibration–in–the–small* assesses the absolute forecast accuracy of predictions at individual levels of $\hat{Y}$. When the calibration curve is linear, this can be summarized by the calibration slope and intercept. A more general approach uses the *loess* nonparametric smoother to estimate the calibration curve[37]. For any shape of calibration curve, errors can be summarized by quantities such as the maximum absolute calibration error, mean absolute calibration error, and 0.9 quantile of calibration error.

The g-index is a new measure of a model's predictive discrimination based only on $X\hat{\beta} = \hat{Y}$ that applies quite generally. It is based on Gini's mean difference for a variable Z, which is the mean over all possible $i \neq j$ of $|Z_i - Z_j|$. The g-index is an interpretable, robust, and highly efficient measure of variation. For example, when predicting systolic blood pressure, $g = 11$mmHg represents a typical difference in $\hat{Y}$. g is independent of censoring and other complexities. For models in which the anti-log of a difference in $\hat{Y}$ represents meaningful ratios (e.g., odds ratios, hazard ratios, ratio of medians), g_r can be defined as $\exp(g)$. For models in which $\hat{Y}$ can be turned into a probability estimate (e.g., logistic regression), g_p is defined as Gini's mean difference of $\hat{P}$. These g–indexes represent e.g. "typical" odds ratios, and "typical" risk differences. Partial g indexes can also be defined. More details may be found in the documentation for the R `rms` package's `gIndex` function.

[a] The s.d. of a binary variable is, aside from a multiplier of $\frac{n}{n-1}$, equal to $\sqrt{a(1-a)}$, where a is the proportion of ones.

[b] There are decompositions of the Brier score into discrimination and calibration components.

Chapter 10
Binary Logistic Regression

10.1 Model

Binary responses are commonly studied in many fields. Examples include the presence or absence of a particular disease, death during surgery, or a consumer purchasing a product. Often one wishes to study how a set of predictor variables X is related to a dichotomous response variable Y. The predictors may describe such quantities as treatment assignment, dosage, risk factors, and calendar time.

For convenience we define the response to be $Y = 0$ or 1, with $Y = 1$ denoting the occurrence of the event of interest. Often a dichotomous outcome can be studied by calculating certain proportions, for example, the proportion of deaths among females and the proportion among males. However, in many situations, there are multiple descriptors, or one or more of the descriptors are continuous. Without a statistical model, studying patterns such as the relationship between age and occurrence of a disease, for example, would require the creation of arbitrary age groups to allow estimation of disease prevalence as a function of age.

Letting X denote the vector of predictors $\{X_1, X_2, \ldots, X_k\}$, a first attempt at modeling the response might use the ordinary linear regression model

$$E\{Y|X\} = X\beta, \tag{10.1}$$

since the expectation of a binary variable Y is $\text{Prob}\{Y = 1\}$. However, such a model by definition cannot fit the data over the whole range of the predictors since a purely linear model $\text{E}\{Y|X\} = \text{Prob}\{Y = 1|X\} = X\beta$ can allow $\text{Prob}\{Y = 1\}$ to exceed 1 or fall below 0. The statistical model that is generally preferred for the analysis of binary responses is instead the binary logistic regression model, stated in terms of the probability that $Y = 1$ given X, the values of the predictors:

F.E. Harrell, Jr., *Regression Modeling Strategies*, Springer Series

$$\text{Prob}\{Y = 1|X\} = [1 + \exp(-X\beta)]^{-1}. \quad (10.2)$$

As before, $X\beta$ stands for $\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$. The binary logistic regression model was developed primarily by Cox[129] and Walker and Duncan.[647] The regression parameters β are estimated by the method of maximum likelihood (see below).

The function

$$P = [1 + \exp(-x)]^{-1} \quad (10.3)$$

is called the logistic function. This function is plotted in Figure 10.1 for x varying from -4 to $+4$. This function has an unlimited range for x while P is restricted to range from 0 to 1.

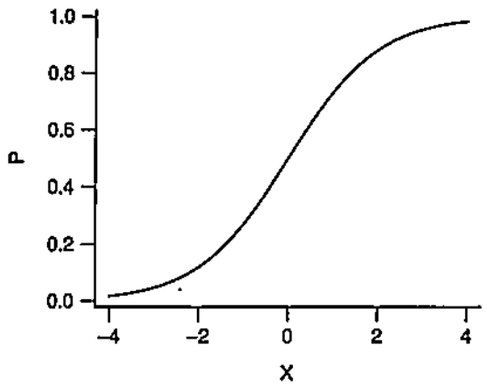


Fig. 10.1 Logistic function

For future derivations it is useful to express x in terms of P. Solving the equation above for x by using

$$1 - P = \exp(-x)/[1 + \exp(-x)] \quad (10.4)$$

yields the inverse of the logistic function:

$$x = \log[P/(1-P)] = \log[\text{odds that } Y = 1 \text{ occurs}] = \text{logit}\{Y = 1\}. \quad (10.5)$$

Other methods that have been used to analyze binary response data include the probit model, which writes P in terms of the cumulative normal distribution, and discriminant analysis. Probit regression, although assuming a similar shape to the logistic function for the regression relationship between $X\beta$ and $\text{Prob}\{Y = 1\}$, involves more cumbersome calculations, and there is no natural interpretation of its regression parameters. In the past, discriminant analysis has been the predominant method since it is the simplest computationally. However, it makes more assumptions than logistic regression. The model used in discriminant analysis is stated in terms of the

distribution of X given the outcome group Y, even though one is seldom interested in the distribution of the predictors per se. The discriminant model has to be inverted using Bayes' rule to derive the quantity of primary interest, $\text{Prob}\{Y = 1\}$. By contrast, the logistic model is a *direct probability model* since it is stated in terms of $\text{Prob}\{Y = 1|X\}$. Since the distribution of a binary random variable Y is completely defined by the true probability that $Y = 1$ and since the model makes no assumption about the distribution of the predictors, the logistic model makes no distributional assumptions whatsoever.

10.1.1 Model Assumptions and Interpretation of Parameters

Since the logistic model is a direct probability model, its only assumptions relate to the form of the regression equation. Regression assumptions are verifiable, unlike the assumption of multivariate normality made by discriminant analysis. The logistic model assumptions are most easily understood by transforming $\text{Prob}\{Y = 1\}$ to make a model that is linear in $X\beta$:

$$\begin{aligned}\text{logit}\{Y = 1|X\} &= \text{logit}(P) = \log[P/(1-P)] \\ &= X\beta, \end{aligned} \quad (10.6)$$

where $P = \text{Prob}\{Y = 1|X\}$. Thus the model is a linear regression model in the log odds that $Y = 1$ since $\text{logit}(P)$ is a weighted sum of the Xs. If all effects are additive (i.e., no interactions are present), the model assumes that for every predictor X_j,

$$\begin{aligned}\text{logit}\{Y = 1|X\} &= \beta_0 + \beta_1 X_1 + \ldots + \beta_j X_j + \ldots + \beta_k X_k \\ &= \beta_j X_j + C, \end{aligned} \quad (10.7)$$

where if all other factors are held constant, C is a constant given by

$$C = \beta_0 + \beta_1 X_1 + \ldots + \beta_{j-1} X_{j-1} + \beta_{j+1} X_{j+1} + \ldots + \beta_k X_k. \quad (10.8)$$

The parameter β_j is then the change in the log odds per unit change in X_j if X_j represents a single factor that is linear and does not interact with other factors and if all other factors are held constant. Instead of writing this relationship in terms of log odds, it could just as easily be written in terms of the odds that $Y = 1$:

$$\text{odds}\{Y = 1|X\} = \exp(X\beta), \quad (10.9)$$

and if all factors other than X_j are held constant,

$$\text{odds}\{Y = 1|X\} = \exp(\beta_j X_j + C) = \exp(\beta_j X_j)\exp(C). \tag{10.10}$$

The regression parameters can also be written in terms of *odds ratios*. The odds that $Y = 1$ when X_j is increased by d, divided by the odds at X_j is

$$\begin{aligned}\frac{\text{odds}\{Y = 1|X_1, X_2, \ldots, X_j + d, \ldots, X_k\}}{\text{odds}\{Y = 1|X_1, X_2, \ldots, X_j, \ldots, X_k\}} \\ = \frac{\exp[\beta_j(X_j + d)]\exp(C)}{[\exp(\beta_j X_j)\exp(C)]} \\ = \exp[\beta_j X_j + \beta_j d - \beta_j X_j] = \exp(\beta_j d).\end{aligned} \tag{10.11}$$

Thus the effect of increasing X_j by d is to increase the odds that $Y = 1$ by a factor of $\exp(\beta_j d)$, or to increase the log odds that $Y = 1$ by an increment of $\beta_j d$. In general, the ratio of the odds of response for an individual with predictor variable values X^* compared with an individual with predictors X is

$$\begin{aligned}X^* : X \text{ odds ratio} &= \exp(X^*\beta)/\exp(X\beta) \\ &= \exp[(X^* - X)\beta].\end{aligned} \tag{10.12}$$

Now consider some special cases of the logistic multiple regression model. If there is only one predictor X and that predictor is binary, the model can be written

$$\begin{aligned}\text{logit}\{Y = 1|X = 0\} &= \beta_0 \\ \text{logit}\{Y = 1|X = 1\} &= \beta_0 + \beta_1.\end{aligned} \tag{10.13}$$

Here β_0 is the log odds of $Y = 1$ when $X = 0$. By subtracting the two equations above, it can be seen that β_1 is the difference in the log odds when $X = 1$ as compared with $X = 0$, which is equivalent to the log of the ratio of the odds when $X = 1$ compared with the odds when $X = 0$. The quantity $\exp(\beta_1)$ is the odds ratio for $X = 1$ compared with $X = 0$. Letting $P^0 = \text{Prob}\{Y = 1|X = 0\}$ and $P^1 = \text{Prob}\{Y = 1|X = 1\}$, the regression parameters are interpreted by

$$\begin{aligned}\beta_0 &= \text{logit}(P^0) = \log[P^0/(1 - P^0)] \\ \beta_1 &= \text{logit}(P^1) - \text{logit}(P^0) \\ &= \log[P^1/(1 - P^1)] - \log[P^0/(1 - P^0)] \\ &= \log\{[P^1/(1 - P^1)]/[P^0/(1 - P^0)]\}.\end{aligned} \tag{10.14}$$

Since there are only two quantities to model and two free parameters, there is no way that this two-sample model can't fit; the model in this case is essentially fitting two cell proportions. Similarly, if there are $g - 1$ dummy indicator Xs representing g groups, the ANOVA-type logistic model must always fit.

If there is one continuous predictor X, the model is

$$\text{logit}\{Y = 1|X\} = \beta_0 + \beta_1 X, \tag{10.15}$$

and without further modification (e.g., taking log transformation of the predictor), the model assumes a straight line in the log odds, or that an increase in X by one unit increases the odds by a factor of $\exp(\beta_1)$.

Now consider the simplest analysis of covariance model in which there are two treatments (indicated by $X_1 = 0$ or 1) and one continuous covariable (X_2). The simplest logistic model for this setup is

$$\text{logit}\{Y = 1|X\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2, \tag{10.16}$$

which can be written also as

$$\begin{aligned}\text{logit}\{Y = 1|X_1 = 0, X_2\} &= \beta_0 + \beta_2 X_2 \\ \text{logit}\{Y = 1|X_1 = 1, X_2\} &= \beta_0 + \beta_1 + \beta_2 X_2.\end{aligned} \tag{10.17}$$

The $X_1 = 1 : X_1 = 0$ odds ratio is $\exp(\beta_1)$, independent of X_2. The odds ratio for a one-unit increase in X_2 is $\exp(\beta_2)$, independent of X_1.

This model, with no term for a possible interaction between treatment and covariable, assumes that for each treatment the relationship between X_2 and log odds is linear, and that the lines have equal slope; that is, they are parallel. Assuming linearity in X_2, the only way that this model can fail is for the two slopes to differ. Thus, the only assumptions that need verification are linearity and lack of interaction between X_1 and X_2.

To adapt the model to allow or test for interaction, we write

$$\text{logit}\{Y = 1|X\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3, \tag{10.18}$$

where the derived variable X_3 is defined to be $X_1 X_2$. The test for lack of interaction (equal slopes) is $H_0 : \beta_3 = 0$. The model can be amplified as

$$\begin{aligned}\text{logit}\{Y = 1|X_1 = 0, X_2\} &= \beta_0 + \beta_2 X_2 \\ \text{logit}\{Y = 1|X_1 = 1, X_2\} &= \beta_0 + \beta_1 + \beta_2 X_2 + \beta_3 X_2 \\ &= \beta_0' + \beta_2' X_2,\end{aligned} \tag{10.19}$$

age group ($< 45, 45 - 54, \geq 55$). The age points on the abscissa for these groups are the overall mean ages in the three age intervals (40.2, 49.1, and 61.1, respectively).

```
require(rms)
```

```
getHdata(sex.age.response)
d ← sex.age.response
dd ← datadist(d); options(datadist='dd')
f ← lrm(response ~ sex + age, data=d)
fasr ← f   # Save for later
w ← function(...)
  with(d, {
    m ← sex=='male'
    f ← sex=='female'
    lpoints(age[f], response[f], pch=1)
    lpoints(age[m], response[m], pch=2)
    af ← cut2(age, c(45,55), levels.mean=TRUE)
    prop ← tapply(response, list(af, sex), mean,
                  na.rm=TRUE)
    agem ← as.numeric(row.names(prop))
    lpoints(agem, prop[,'female'],
            pch=4, cex=1.3, col='green')
    lpoints(agem, prop[,'male'],
            pch=5, cex=1.3, col='green')
    x ← rep(62, 4); y ← seq(.25, .1, length=4)
    lpoints(x, y, pch=c(1, 2, 4, 5),
            col=rep(c('blue','green'),each=2))
    ltext(x+5, y,
          c('F Observed','M Observed',
            'F Proportion','M Proportion'), cex=.8)
  } )   # Figure 10.3

plot(Predict(f, age=seq(34, 70, length=200), sex, fun=plogis),
     ylab='Pr[response]', ylim=c(-.02, 1.02), addpanel=w)
ltx ← function(fit) latex(fit, inline=TRUE, columns=54,
                          file='', after='$.', digits=3,
        size='Ssize', before='$X\\hat{\\beta}=')
ltx(f)
```

$X\hat{\beta} = -9.84 + 3.49[\text{male}] + 0.158\,\text{age}$.

Descriptive statistics for assessing the association between sex and response, age group and response, and age group and response stratified by sex are found below. Corresponding fitted logistic models, with sex coded as 0 = female, 1 = male are also given. Models were fitted first with sex as the only predictor, then with age as the (continuous) predictor, then with sex and age simultaneously. First consider the relationship between sex and response, ignoring the effect of age.

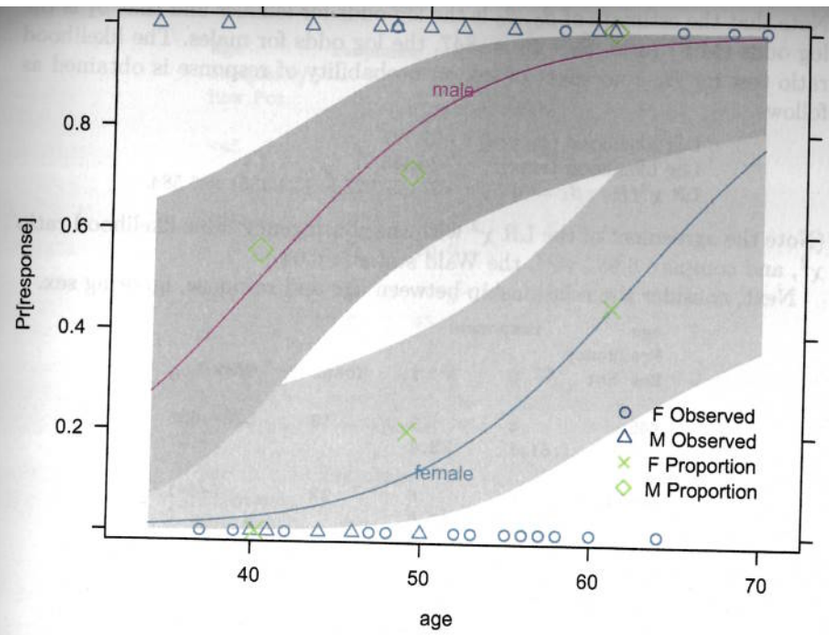


Fig. 10.3 Data, subgroup proportions, and fitted logistic model, with 0.95 pointwise confidence bands

sex	response			
Frequency				
Row Pct	0	1	Total	Odds/Log
F	14	6	20	6/14=.429
	70.00	30.00		-.847
M	6	14	20	14/6=2.33
	30.00	70.00		.847
Total	20	20	40	

M:F odds ratio = (14/6)/(6/14) = 5.44, log=1.695

Statistics for sex × response

Statistic	d.f.	Value	P
χ^2	1	6.400	0.011
Likelihood Ratio χ^2	1	6.583	0.010

Parameter	Estimate	Std Err	Wald χ^2	P
β_0	−0.8473	0.4880	3.0152	
β_1	1.6946	0.6901	6.0305	0.0141

Note that the estimate of β_0, $\hat{\beta}_0$ is the log odds for females and that $\hat{\beta}_1$ is the log odds (M:F) ratio. $\hat{\beta}_0 + \hat{\beta}_1 = .847$, the log odds for males. The likelihood ratio test for H_0 : no effect of sex on probability of response is obtained as follows.

Log likelihood ($\beta_1 = 0$) : -27.727
Log likelihood (max) : -24.435
LR $\chi^2(H_0 : \beta_1 = 0)$: $-2(-27.727 - -24.435) = 6.584$.

(Note the agreement of the LR χ^2 with the contingency table likelihood ratio χ^2, and compare 6.584 with the Wald statistic 6.03.)

Next, consider the relationship between age and response, ignoring sex.

age / Frequency / Row Pct	response 0	1	Total	Odds/Log
<45	8 61.5	5 38.4	13	5/8=.625 -.47
45-54	6 50.0	6 50.0	12	6/6=1 0
55+	6 40.0	9 60.0	15	9/6=1.5 .405
Total	20	20	40	

55+ : <45 odds ratio = (9/6)/(5/8) = 2.4, log=.875

Parameter	Estimate	Std Err	Wald χ^2	P
β_0	−2.7338	1.8375	2.2134	0.1368
β_1	0.0540	0.0358	2.2763	0.1314

The estimate of β_1 is in rough agreement with that obtained from the frequency table. The 55+ : < 45 log odds ratio is .875, and since the respective mean ages in the 55+ and <45 age groups are 61.1 and 40.2, an estimate of the log odds ratio increase per year is $.875/(61.1 - 40.2) = .875/20.9 = .042$.

The likelihood ratio test for H_0 : no association between age and response is obtained as follows.

Log likelihood ($\beta_1 = 0$) : -27.727
Log likelihood (max) : -26.511
LR $\chi^2(H_0 : \beta_1 = 0)$: $-2(-27.727 - -26.511) = 2.432$.

(Compare 2.432 with the Wald statistic 2.28.)

Next we consider the simultaneous association of age and sex with response.

sex=F

age / Frequency / Row Pct	response 0	1	Total
<45	4 100.0	0 0.0	4
45-54	4 80.0	1 20.0	5
55+	6 54.6	5 45.4	11
Total	14	6	20

sex=M

age / Frequency / Row Pct	response 0	1	Total
<45	4 44.4	5 55.6	9
45-54	2 28.6	5 71.4	7
55+	0 0.0	4 100.0	4
Total	6	14	20

A logistic model for relating sex and age simultaneously to response is given below.

Parameter	Estimate	Std Err	Wald χ^2	P
β_0	−9.8429	3.6758	7.1706	0.0074
β_1 (sex)	3.4898	1.1992	8.4693	0.0036
β_2 (age)	0.1581	0.0616	6.5756	0.0103

Likelihood ratio tests are obtained from the information below.

Log likelihood ($\beta_1 = 0, \beta_2 = 0$) : -27.727
Log likelihood (max) : -19.458
Log likelihood ($\beta_1 = 0$) : -26.511
Log likelihood ($\beta_2 = 0$) : -24.435
LR χ^2 ($H_0 : \beta_1 = \beta_2 = 0$) : $-2(-27.727 - -19.458) = 16.538$
LR χ^2 ($H_0 : \beta_1 = 0$) sex|age : $-2(-26.511 - -19.458) = 14.106$
LR χ^2 ($H_0 : \beta_2 = 0$) age|sex : $-2(-24.435 - -19.458) = 9.954$.

The 14.1 should be compared with the Wald statistic of 8.47, and 9.954 should be compared with 6.58. The fitted logistic model is plotted separately

for females and males in Figure 10.3. The fitted model is

$$\text{logit}\{\text{Response} = 1|\text{sex,age}\} = -9.84 + 3.49 \times \text{sex} + .158 \times \text{age}, \quad (10.21)$$

where as before sex = 0 for females, 1 for males. For example, for a 40-year-old female, the predicted logit is $-9.84 + .158(40) = -3.52$. The predicted probability of a response is $1/[1 + \exp(3.52)] = .029$. For a 40-year-old male, the predicted logit is $-9.84 + 3.49 + .158(40) = -.03$, with a probability of .492.

10.1.4 Design Formulations

The logistic multiple regression model can incorporate the same designs as can ordinary linear regression. An analysis of variance (ANOVA) model for a treatment with k levels can be formulated with $k - 1$ dummy variables. This logistic model is equivalent to a $2 \times k$ contingency table. An analysis of covariance logistic model is simply an ANOVA model augmented with covariables used for adjustment.

One unique design that is interesting to consider in the context of logistic models is a simultaneous comparison of multiple factors between two groups. Suppose, for example, that in a randomized trial with two treatments one wished to test whether any of 10 baseline characteristics are mal-distributed between the two groups. If the 10 factors are continuous, one could perform a two-sample Wilcoxon–Mann–Whitney test or a t-test for each factor (if each is normally distributed). However, this procedure would result in multiple comparison problems and would also not be able to detect the combined effect of small differences across all the factors. A better procedure would be a multivariate test. The Hotelling T^2 test is designed for just this situation. It is a k-variable extension of the one-variable unpaired t-test. The T^2 test, like discriminant analysis, assumes multivariate normality of the k factors. This assumption is especially tenuous when some of the factors are polytomous. A better alternative is the global test of no regression from the logistic model. This test is valid because it can be shown that H_0 : mean X is the same for both groups (= H_0 : mean X does not depend on group = H_0 : mean $X|$ group = constant) is true if and only if H_0 : Prob{group$|X$} = constant. Thus k factors can be tested simultaneously for differences between the two groups using the binary logistic model, which has far fewer assumptions than does the Hotelling T^2 test. The logistic global test of no regression (with k d.f.) would be expected to have greater power if there is non-normality. Since the logistic model makes no assumption regarding the distribution of the descriptor variables, it can easily test for simultaneous group differences involving a mixture of continuous, binary, and nominal variables. In observational studies, such models for treatment received or exposure (propensity score models) hold great promise for adjusting for confounding.[117,380,526,530,531]

O'Brien[479] has developed a general test for comparing group 1 with group 2 for a single measurement. His test detects location and scale differences by fitting a logistic model for Prob{Group 2} using X and X^2 as predictors.

For a randomized study where adjustment for confounding is seldom necessary, adjusting for covariables using a binary logistic model results in *increases* in standard errors of regression coefficients.[527] This is the opposite of what happens in linear regression where there is an unknown variance parameter that is estimated using the residual squared error. Fortunately, adjusting for covariables using logistic regression, by accounting for subject heterogeneity, will result in larger regression coefficients even for a randomized treatment variable. The increase in estimated regression coefficients more than offsets the increase in standard error[284,285,527,588].

10.2 Estimation

10.2.1 Maximum Likelihood Estimates

The parameters in the logistic regression model are estimated using the maximum likelihood (ML) method. The method is based on the same principles as the one-sample proportion example described in Section 9.1. The difference is that the general logistic model is not a single sample or a two-sample problem. The probability of response for the ith subject depends on a particular set of predictors X_i, and in fact the list of predictors may not be the same for any two subjects. Denoting the response and probability of response of the ith subject by Y_i and P_i, respectively, the model states that

$$P_i = \text{Prob}\{Y_i = 1|X_i\} = [1 + \exp(-X_i\beta)]^{-1}. \quad (10.22)$$

The likelihood of an observed response Y_i given predictors X_i and the unknown parameters β is

$$P_i^{Y_i}[1 - P_i]^{1-Y_i}. \quad (10.23)$$

The joint likelihood of all responses $Y_1, Y_2, \ldots, Y_n$ is the product of these likelihoods for $i = 1, \ldots, n$. The likelihood and log likelihood functions are rewritten by using the definition of P_i above to allow them to be recognized as a function of the unknown parameters β. Except in simple special cases (such as the k-sample problem in which all Xs are dummy variables), the ML estimates (MLE) of β cannot be written explicitly. The Newton–Raphson method described in Section 9.4 is usually used to solve iteratively for the list of values β that maximize the log likelihood. The MLEs are denoted by

estimate of the variance–covariance matrix of β.

Under $H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0$, the intercept parameter β_0 can be estimated explicitly and the log likelihood under this global null hypothesis can be computed explicitly. Under the global null hypothesis, $P_i = P = [1 + \exp(-\beta_0)]^{-1}$ and the MLE of P is $\hat{P} = s/n$ where s is the number of responses and n is the sample size. The MLE of β_0 is $\hat{\beta}_0 = \text{logit}(\hat{P})$. The log likelihood under this null hypothesis is

$$\begin{aligned} & s \log(\hat{P}) + (n-s)\log(1-\hat{P}) \\ = \; & s \log(s/n) + (n-s)\log[(n-s)/n] \\ = \; & s \log s + (n-s)\log(n-s) - n\log(n). \end{aligned} \tag{10.24}$$

10.2.2 Estimation of Odds Ratios and Probabilities

Once β is estimated, one can estimate any log odds, odds, or odds ratios. The MLE of the $X_j + 1 : X_j$ log odds ratio is $\hat{\beta}_j$, and the estimate of the $X_j + d : X_j$ log odds ratio is $\hat{\beta}_j d$, all other predictors remaining constant (assuming the absence of interactions and nonlinearities involving X_j). For large enough samples, the MLEs are normally distributed with variances that are consistently estimated from the estimated variance–covariance matrix. Letting z denote the $1-\alpha/2$ critical value of the standard normal distribution, a two-sided $1-\alpha$ confidence interval for the log odds ratio for a one-unit increase in X_j is $[\hat{\beta}_j - zs, \hat{\beta}_j + zs]$, where s is the estimated standard error of $\hat{\beta}_j$. (Note that for $\alpha = .05$, i.e., for a 95% confidence interval, $z = 1.96$.)

A theorem in statistics states that the MLE of a function of a parameter is that same function of the MLE of the parameter. Thus the MLE of the $X_j + 1 : X_j$ odds ratio is $\exp(\hat{\beta}_j)$. Also, if a $1-\alpha$ confidence interval of a parameter β is $[c, d]$ and $f(u)$ is a one-to-one function, a $1-\alpha$ confidence interval of $f(\beta)$ is $[f(c), f(d)]$. Thus a $1-\alpha$ confidence interval for the $X_j+1 : X_j$ odds ratio is $\exp[\hat{\beta}_j \pm zs]$. Note that while the confidence interval for β_j is symmetric about $\hat{\beta}_j$, the confidence interval for $\exp(\beta_j)$ is not. By the same theorem just used, the MLE of $P_i = \text{Prob}\{Y_i = 1|X_i\}$ is

$$\hat{P}_i = [1 + \exp(-X_i\hat{\beta})]^{-1}. \tag{10.25}$$

A confidence interval for P_i could be derived by computing the standard error of $\hat{P}_i$, yielding a symmetric confidence interval. However, such an interval would have the disadvantage that its endpoints could fall below zero or exceed one. A better approach uses the fact that for large samples $X\hat{\beta}$ is approximately normally distributed. An estimate of the variance of $X\hat{\beta}$ in matrix notation is XVX' where V is the estimated variance–covariance covariances of $\hat{\beta}$ weighted by squares and products of the predictors. The estimated standard error of $X\hat{\beta}$, s, is the square root of this variance estimate. A $1-\alpha$ confidence interval for P_i is then

$$\{1 + \exp[-(X_i\hat{\beta} \pm zs)]\}^{-1}. \tag{10.26}$$

10.2.3 Minimum Sample Size Requirement

Suppose there were no covariates, so that the only parameter in the model is the intercept. What is the sample size required to allow the estimate of the intercept to be precise enough so that the predicted probability is within 0.1 of the true probability with 0.95 confidence, when the true intercept is in the neighborhood of zero? The answer is n=96. What if there were one covariate, and it was binary with a prevalence of $\frac{1}{2}$? One would need 96 subjects with $X = 0$ and 96 with $X = 1$ to have an upper bound on the margin of error for estimating $\text{Prob}\{Y = 1|X = x\}$ not exceed 0.1 for either value of x[a].

Now consider a very simple single continuous predictor case in which X has a normal distribution with mean zero and standard deviation σ, with the true $\text{Prob}\{Y = 1|X = x\} = [1 + \exp(-x)]^{-1}$. The expected number of events is $\frac{n}{2}$[b]. The following simulation answers the question "What should n be so that the expected maximum absolute error (over $x \in [-1.5, 1.5]$) in $\hat{P}$ is less than ϵ?"

```
sigmas  ← c(.5, .75, 1, 1.25, 1.5, 1.75, 2, 2.5, 3, 4)
ns      ← seq(25, 300, by=25)
nsim    ← 1000
xs      ← seq(-1.5, 1.5, length=200)
pactual ← plogis(xs)

dn ← list(sigma=format(sigmas), n=format(ns))
maxerr ← N1 ← array(NA, c(length(sigmas), length(ns)), dn)
require(rms)

i ← 0
for(s in sigmas) {
  i ← i + 1
  j ← 0
  for(n in ns) {
```

[a] The general formula for the sample size required to achieve a margin of error of δ in estimating a true probability of θ at the 0.95 confidence level is $n = (\frac{1.96}{\delta})^2 \times \theta(1-\theta)$. Set $\theta = \frac{1}{2}$ (intercept=0) for the worst case.

[b] The R code can easily be modified for other event frequencies, or the minimum of the number of events and non-events for a dataset at hand can be compared with $\frac{n}{2}$ in this simulation. An average maximum absolute error of 0.05 corresponds roughly to a half-width of the 0.95 confidence interval of 0.1.

```
    n1 ← maxe ← 0
    for(k in 1:nsim) {
      x ← rnorm(n, 0, s)
      P ← plogis(x)
      y ← ifelse(runif(n) ≤ P, 1, 0)
      n1 ← n1 + sum(y)
      beta ← lrm.fit(x, y)$coefficients
      phat ← plogis(beta[1] + beta[2] * xs)
      maxe ← maxe + max(abs(phat - pactual))
    }
    n1 ← n1/nsim
    maxe ← maxe/nsim
    maxerr[i,j] ← maxe
    N1[i,j] ← n1
  }
}
xrange ← range(xs)
simerr ← llist(N1, maxerr, sigmas, ns, nsim, xrange)

maxe ← reShape(maxerr)
# Figure 10.4
xYplot(maxerr ~ n, groups=sigma, data=maxe,
       ylab=expression(paste('Average Maximum  ',
            abs(hat(P) - P))),
       type='l', lty=rep(1:2, 5), label.curve=FALSE,
       abline=list(h=c(.15, .1, .05), col=gray(.85)))
Key(.8, .68, other=list(cex=.7,
                title=expression(~~~~~~~~~~sigma)))
```

10.3 Test Statistics

The likelihood ratio, score, and Wald statistics discussed earlier can be used to test any hypothesis in the logistic model. The likelihood ratio test is generally preferred. When true parameters are near the null values all three statistics usually agree. The Wald test has a significant drawback when the true parameter value is very far from the null value. In such case the standard error estimate becomes too large. As $\hat{\beta}_j$ increases from 0, the Wald test statistic for $H_0 : \beta_j = 0$ becomes larger, but after a certain point it becomes smaller. The statistic will eventually drop to zero if $\hat{\beta}_j$ becomes infinite.[278] Infinite estimates can occur in the logistic model especially when there is a binary predictor whose mean is near 0 or 1. Wald statistics are especially problematic in this case. For example, if 10 out of 20 males had a disease and 5 out of 5 females had the disease, the female : male odds ratio is infinite and so is the logistic regression coefficient for sex. If such a situation occurs, the likelihood ratio or score statistic should be used instead of the Wald statistic.

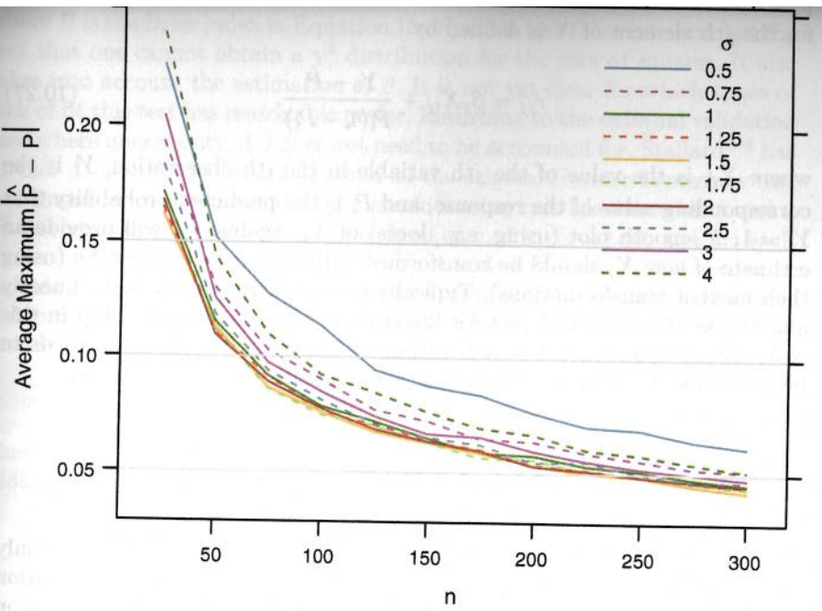


Fig. 10.4 Simulated expected maximum error in estimating probabilities for $x \in [-1.5, 1.5]$ with a single normally distributed X with mean zero

For k-sample (ANOVA-type) logistic models, logistic model statistics are equivalent to contingency table χ^2 statistics. As exemplified in the logistic model relating sex to response described previously, the global likelihood ratio statistic for all dummy variables in a k-sample model is identical to the contingency table (k-sample binomial) likelihood ratio χ^2 statistic. The score statistic for this same situation turns out to be identical to the $k-1$ degrees of freedom Pearson χ^2 for a $k \times 2$ table.

As mentioned in Section 2.6, it can be dangerous to interpret individual parameters, make pairwise treatment comparisons, or test linearity if the overall test of association for a factor represented by multiple parameters is insignificant.

10.4 Residuals

Several types of residuals can be computed for binary logistic model fits. Many of these residuals are used to examine the influence of individual observations on the fit. The *partial residual* can be used for directly assessing how each

predictor should be transformed for the with-adequate-[illegible]

for the jth element of X is defined by

$$r_{ij} = \hat{\beta}_j X_{ij} + \frac{Y_i - \hat{P}_i}{\hat{P}_i(1 - \hat{P}_i)}, \tag{10.27}$$

where X_{ij} is the value of the jth variable in the ith observation, Y_i is the corresponding value of the response, and $\hat{P}_i$ is the predicted probability that $Y_i = 1$. A smooth plot (using, e.g., loess) of X_{ij} against r_{ij} will provide an estimate of how X_j should be transformed, adjusting for the other Xs (using their current transformations). Typically one tentatively models X_j linearly and checks the smoothed plot for linearity. A U-shaped relationship in this plot, for example, indicates that a squared term or spline function needs to be added for X_j. This approach does assume additivity of predictors.

9

10.5 Assessment of Model Fit

As the logistic regression model makes no distributional assumptions, only the assumptions of linearity and additivity need to be verified (in addition to the usual assumptions about independence of observations and inclusion of important covariables). In ordinary linear regression there is no global test for lack of model fit unless there are replicate observations at various settings of X. This is because ordinary regression entails estimation of a separate variance parameter σ^2. In logistic regression there are global tests for goodness of fit. Unfortunately, some of the most frequently used ones are inappropriate. For example, it is common to see a deviance test of goodness of fit based on the "residual" log likelihood, with P-values obtained from a χ^2 distribution with $n - p$ d.f. This P-value is inappropriate since the deviance does not have an asymptotic χ^2 distribution, due to the facts that the number of parameters estimated is increasing at the same rate as n and the expected cell frequencies are far below five (by definition).

Hosmer and Lemeshow[304] have developed a commonly used test for goodness of fit for binary logistic models based on grouping into deciles of predicted probability and performing an ordinary χ^2 test for the mean predicted probability against the observed fraction of events (using 8 d.f. to account for evaluating fit on the model development sample). The Hosmer–Lemeshow test is dependent on the choice of how predictions are grouped[303] and it is not clear that the choice of the number of groups should be independent of n. Hosmer et al.[303] have compared a number of global goodness of fit tests for binary logistic regression. They concluded that the simple unweighted sum of squares test of Copas[124] as modified by le Cessie and van Houwelingen[387] is as

[illegible] where B is the Brier index in Equation 10.35). This test takes into account the fact that one cannot obtain a χ^2 distribution for the sum of squares. It also takes into account the estimation of β. It is not yet clear for which types of lack of fit this test has reasonable power. Returning to the external validation case where uncertainty of β does not need to be accounted for, Stallard[584] has further documented the lack of power of the original Hosmer-Lemeshow test and found more power with a logarithmic scoring rule (deviance test) and a χ^2 test that, unlike the simple unweighted sum of squares test, weights each squared error by dividing it by $\hat{P}_i(1 - \hat{P}_i)$. A scaled χ^2 distribution seemed to provide the best approximation to the null distribution of the test statistics.

More power for detecting lack of fit is expected to be obtained from testing specific alternatives to the model. In the model

$$\text{logit}\{Y = 1|X\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2, \tag{10.28}$$

where X_1 is binary and X_2 is continuous, one needs to verify that the log odds is related to X_1 and X_2 according to Figure 10.5.

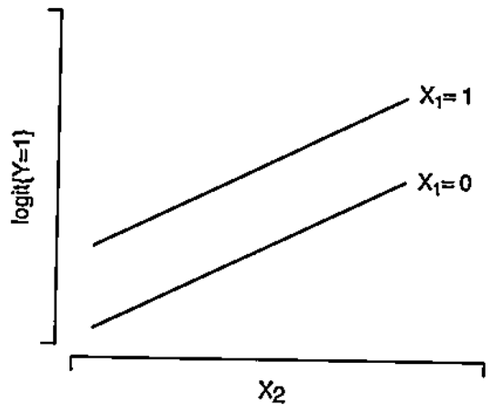


Fig. 10.5 Logistic regression assumptions for one binary and one continuous predictor

The simplest method for validating that the data are consistent with the no-interaction linear model involves stratifying the sample by X_1 and quantile groups (e.g., deciles) of X_2.[265] Within each stratum the proportion of responses $\hat{P}$ is computed and the log odds calculated from $\log[\hat{P}/(1 - \hat{P})]$. The number of quantile groups should be such that there are at least 20 (and perhaps many more) subjects in each $X_1 \times X_2$ group. Otherwise, probabilities cannot be estimated precisely enough to allow trends to be seen above "noise" in the data. Since at least 3 X_2 groups must be formed to allow assessment of linearity, the total sample size must be at least $2 \times 3 \times 20 = 120$ for this method to work at all.

Figure 10.6 demonstrates this method for a large sample size of 3504 subjects stratified by sex and deciles of age. Linearity is apparent for males while there is evidence for slight interaction between age and sex since the age trend for females appears curved.

```
getHdata(acath)
acath$sex ← factor(acath$sex, 0:1, c('male','female'))
dd ← datadist(acath); options(datadist='dd')
f ← lrm(sigdz ~ rcs(age, 4) * sex, data=acath)
```

```
w ← function(...)
  with(acath, {
    plsmo(age, sigdz, group=sex, fun=qlogis, lty='dotted',
          add=TRUE, grid=TRUE)
    af ← cut2(age, g=10, levels.mean=TRUE)
    prop ← qlogis(tapply(sigdz, list(af, sex), mean,
                         na.rm=TRUE))
    agem ← as.numeric(row.names(prop))
    lpoints(agem, prop[,'female'], pch=4, col='green')
    lpoints(agem, prop[,'male'],   pch=2, col='green')
  } )    # Figure 10.6
plot(Predict(f, age, sex), ylim=c(-2,4), addpanel=w,
     label.curve=list(offset=unit(0.5, 'cm')))
```

The subgrouping method requires relatively large sample sizes and does not use continuous factors effectively. The ordering of values is not used at all between intervals, and the estimate of the relationship for a continuous variable has little resolution. Also, the method of grouping chosen (e.g., deciles vs. quintiles vs. rounding) can alter the shape of the plot.

In this dataset with only two variables, it is efficient to use a nonparametric smoother for age, separately for males and females. Nonparametric smoothers, such as `loess`[111] used here, work well for binary response variables (see Section 2.4.7); the logit transformation is made on the smoothed probability estimates. The smoothed estimates are shown in Figure 10.6.

When there are several predictors, the restricted cubic spline function is better for estimating the true relationship between X_2 and $\text{logit}\{Y = 1\}$ for continuous variables without assuming linearity. By fitting a model containing X_2 expanded into $k-1$ terms, where k is the number of knots, one can obtain an estimate of the transformation of X_2 as discussed in Section 2.4:

$$\begin{aligned}\text{logit}\{Y = 1|X\} &= \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_2' + \hat{\beta}_4 X_2'' \\ &= \hat{\beta}_0 + \hat{\beta}_1 X_1 + f(X_2), \end{aligned} \tag{10.29}$$

where X_2' and X_2'' are constructed spline variables (when $k = 4$). Plotting the estimated spline function $f(X_2)$ versus X_2 will estimate how the effect of X_2 should be modeled. If the sample is sufficiently large, the spline function can be fitted separately for $X_1 = 0$ and $X_1 = 1$, allowing detection of even unusual interaction patterns. A formal test of linearity in X_2 is obtained by testing $H_0 : \beta_3 = \beta_4 = 0$.

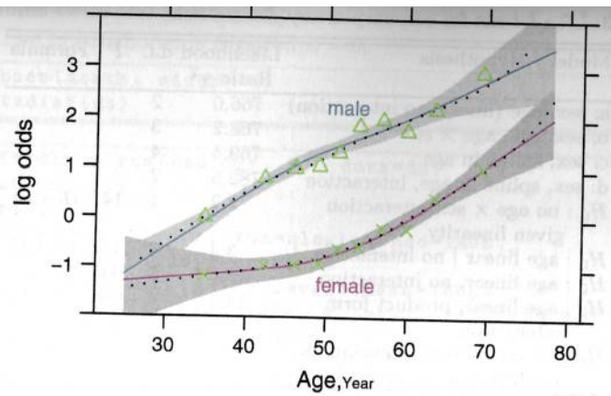


Fig. 10.6 Logit proportions of significant coronary artery disease by sex and deciles of age for n=3504 patients, with spline fits (smooth curves). Spline fits are for $k = 4$ knots at age= 36, 48, 56, and 68 years, and interaction between age and sex is allowed. Shaded bands are pointwise 0.95 confidence limits for predicted log odds. Smooth nonparametric estimates are shown as dotted curves. Data courtesy of the Duke Cardiovascular Disease Databank.

For testing interaction between X_1 and X_2, a product term (e.g., X_1X_2) can be added to the model and its coefficient tested. A more general simultaneous test of linearity and lack of interaction for a two-variable model in which one variable is binary (or is assumed linear) is obtained by fitting the model

$$\begin{aligned}\text{logit}\{Y = 1|X\} = \beta_0 &+ \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_2' + \beta_4 X_2'' \\ &+ \beta_5 X_1 X_2 + \beta_6 X_1 X_2' + \beta_7 X_1 X_2'' \end{aligned} \tag{10.30}$$

and testing $H_0 : \beta_3 = \ldots = \beta_7 = 0$. This formulation allows the shape of the X_2 effect to be completely different for each level of X_1. There is virtually no departure from linearity and additivity that cannot be detected from this expanded model formulation. The most computationally efficient test for lack of fit is the score test (e.g., X_1 and X_2 are forced into a tentative model and the remaining variables are candidates). Figure 10.6 also depicts a fitted spline logistic model with $k = 4$, allowing for general interaction between age and sex as parameterized above. The fitted function, after expanding the restricted cubic spline function for simplicity (see Equation 2.27), is given above. Note the good agreement between the empirical estimates of log odds and the spline fits and nonparametric estimates in this large dataset.

An analysis of log likelihood for this model and various sub-models is found in Table 10.3. The χ^2 for global tests is corrected for the intercept and the degrees of freedom does not include the intercept.

Table 10.3 LR χ^2 tests for coronary artery disease risk

Model / Hypothesis	Likelihood Ratio χ^2	d.f.	P	Formula
a: sex, age (linear, no interaction)	766.0	2		
b: sex, age, age × sex	768.2	3		
c: sex, spline in age	769.4	4		
d: sex, spline in age, interaction	782.5	7		
H_0 : no age × sex interaction given linearity	2.2	1	.14	$(b-a)$
H_0 : age linear | no interaction	3.4	2	.18	$(c-a)$
H_0 : age linear, no interaction	16.6	5	.005	$(d-a)$
H_0 : age linear, product form interaction	14.4	4	.006	$(d-b)$
H_0 : no interaction, allowing for nonlinearity in age	13.1	3	.004	$(d-c)$

Table 10.4 AIC on χ^2 scale by number of knots

k	Model χ^2	AIC
0	99.23	97.23
3	112.69	108.69
4	121.30	115.30
5	123.51	115.51
6	124.41	114.51

This analysis confirms the first impression from the graph, namely, that age × sex interaction is present but it is not of the form of a simple product between age and sex (change in slope). In the context of a linear age effect, there is no significant product interaction effect ($P = .14$). Without allowing for interaction, there is no significant nonlinear effect of age ($P = .18$). However, the general test of lack of fit with 5 d.f. indicates a significant departure from the linear additive model ($P = .005$).

In Figure 10.7, data from 2332 patients who underwent cardiac catheterization at Duke University Medical Center and were found to have significant ($\geq 75\%$) diameter narrowing of at least one major coronary artery were analyzed (the dataset is available from the Web site). The relationship between the time from the onset of symptoms of coronary artery disease (e.g., angina, myocardial infarction) to the probability that the patient has severe (three-vessel disease or left main disease—`tvdlm`) coronary disease was of interest. There were 1129 patients with `tvdlm`. A logistic model was used with the duration of symptoms appearing as a restricted cubic spline function with $k = 3, 4, 5$, and 6 equally spaced knots in terms of quantiles between .05 and .95. The best fit for the number of parameters was chosen using Akaike's information criterion (AIC), computed in Table 10.4 as the model likelihood ratio χ^2 minus twice the number of parameters in the model. intercept. The linear model is denoted $k = 0$.

```
dz ← subset(acath, sigdz==1)
dd ← datadist(dz)
```

```
f ← lrm(tvdlm ~ rcs(cad.dur, 5), data=dz)
w ← function(...)
  with(dz, {
    plsmo(cad.dur, tvdlm, fun=qlogis, add=TRUE,
          grid=TRUE, lty='dotted')
    x ← cut2(cad.dur, g=15, levels.mean=TRUE)
    prop ← qlogis(tapply(tvdlm, x, mean, na.rm=TRUE))
    xm ← as.numeric(names(prop))
    lpoints(xm, prop, pch=2, col='green')
  } )   # Figure 10.7
plot(Predict(f, cad.dur), addpanel=w)
```

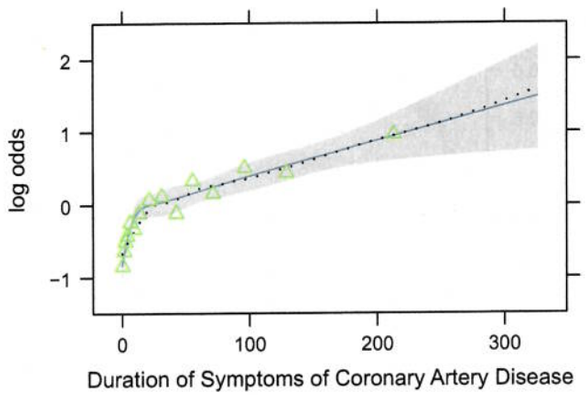


Fig. 10.7 Estimated relationship between duration of symptoms and the log odds of severe coronary artery disease for $k = 5$. Knots are marked with arrows. Solid line is spline fit; dotted line is a nonparametric loess estimate.

Figure 10.7 displays the spline fit for $k = 5$. The triangles represent subgroup estimates obtained by dividing the sample into groups of 150 patients. For example, the leftmost triangle represents the logit of the proportion of `tvdlm` in the 150 patients with the shortest duration of symptoms, versus the mean duration in that group. A Wald test of linearity, with 3 d.f., showed highly significant nonlinearity (χ^2= 23.92 with 3 d.f.). The plot of the spline transformation suggests a log transformation, and when log (duration of symptoms in months + 1) was fitted in a logistic model, the log likelihood of the model (119.33 with 1 d.f.) was virtually as good as the spline model (123.51 with 4 d.f.); the corresponding Akaike information criteria (on the χ^2 scale) are 117.33 and 115.51. To check for adequacy in the log transformation,

d. Plot the estimated logit response as a function of age and sex, with and without fitting an interaction term.
e. Perform a likelihood ratio test of H_0 : the model containing only age and sex is adequate versus H_a : model is inadequate. Here, "inadequate" may mean nonlinearity (quadratic) in age or presence of an interaction.
f. Assuming no interaction is present, test H_0 : model is linear in age versus H_a : model is nonlinear in age. Allow "nonlinear" to be more general than quadratic. (Hint: use a restricted cubic spline function with knots at age=39, 45, 55, 64 years.)
g. Plot age against the estimated spline transformation of age (the transformation that would make age fit linearly). You can set the sex and intercept terms to anything you choose. Also plot Prob{response = 1 | age, sex} from this fitted restricted cubic spline logistic model.

2. Consider a binary logistic regression model using the following predictors: age (years), sex, race (white, African-American, Hispanic, Oriental, other), blood pressure (mmHg). The fitted model is given by

$$\begin{aligned}\text{logit Prob}[Y = 1|X] = X\hat{\beta} = &-1.36 + .03(\text{race} = \text{African-American}) \\ &- .04(\text{race} = \text{hispanic}) + .05(\text{race} = \text{oriental}) - .06(\text{race} = \text{other}) \\ &+ .07|\text{blood pressure} - 110| + .3(\text{sex} = \text{male}) - .1\text{age} + .002\text{age}^2 + \\ &(\text{sex} = \text{male})[.05\text{age} - .003\text{age}^2].\end{aligned}$$

a. Compute the predicted logit (log odds) that $Y = 1$ for a 50-year-old female Hispanic with a blood pressure of 90 mmHg. Also compute the odds that $Y = 1$ (Prob$[Y = 1]$/Prob$[Y = 0]$) and the estimated probability that $Y = 1$.
b. Estimate odds ratios for each nonwhite race compared with the reference group (white), holding all other predictors constant. Why can you estimate the relative effect of race for all types of subjects without specifying their characteristics?
c. Compute the odds ratio for a blood pressure of 120 mmHg compared with a blood pressure of 105, holding age first to 30 years and then to 40 years.
d. Compute the odds ratio for a blood pressure of 120 mmHg compared with a blood pressure of 105, all other variables held to unspecified constants. Why is this relative effect meaningful without knowing the subject's age, race, or sex?
e. Compute the estimated risk difference in changing blood pressure from 105 mmHg to 120 mmHg, first for age = 30 then for age = 40, for a white female. Why does the risk difference depend on age?
f. Compute the relative odds for males compared with females, for age = 50 and other variables held constant.
g. Same as the previous question but for females : males instead of males : females.
h. Compute the odds ratio resulting from increasing age from 50 to 55 for males, and then for females, other variables held constant. What is wrong with the following question: What is the relative effect of chang-

Chapter 11
Case Study in Binary Logistic Regression, Model Selection and Approximation: Predicting Cause of Death

11.1 Overview

This chapter contains a case study on developing, describing, and validating a binary logistic regression model. In addition, the following methods are exemplified:

1. Data reduction using incomplete linear and nonlinear principal components
2. Use of AIC to choose from five modeling variations, deciding which is best for the number of parameters
3. Model simplification using stepwise variable selection and approximation of the full model
4. The relationship between the degree of approximation and the degree of predictive discrimination loss
5. Bootstrap validation that includes penalization for model uncertainty (variable selection) and that demonstrates a loss of predictive discrimination over the full model even when compensating for overfitting the full model.

The data reduction and pre-transformation methods used here were discussed in more detail in Chapter 8. Single imputation will be used because of the limited quantity of missing data.

11.2 Background

Consider the randomized trial of estrogen for treatment of prostate cancer[87] described in Chapter 8. In this trial, larger doses of estrogen reduced the effect of prostate cancer but at the cost of increased risk of cardiovascular death.

F. E. Harrell, Jr., *Regression Modeling Strategies*, Springer Series

Kay[340] did a formal analysis of the competing risks for cancer, cardiovascular, and other deaths. It can also be quite informative to study how treatment and baseline variables relate to the cause of death for those patients who died.[376] We subset the original dataset of those patients dying from prostate cancer ($n = 130$), heart or vascular disease ($n = 96$), or cerebrovascular disease ($n = 31$). Our goal is to predict cardiovascular–cerebrovascular death (cvd, $n = 127$) given the patient died from either cvd or prostate cancer. Of interest is whether the time to death has an effect on the cause of death, and whether the importance of certain variables depends on the time of death.

11.3 Data Transformations and Single Imputation

In R, first obtain the desired subset of the data and do some preliminary calculations such as combining an infrequent category with the next category, and dichotomizing ekg for use in ordinary principal components (PCs).

```
require(rms)
```

```
getHdata(prostate)
prostate ←
  within(prostate, {
    levels(ekg)[levels(ekg) %in%
                c('old MI','recent MI')] ← 'MI'
    ekg.norm ← 1*(ekg %in% c('normal','benign'))
    levels(ekg) ← abbreviate(levels(ekg))
    pfn ← as.numeric(pf)
    levels(pf)  ← levels(pf)[c(1,2,3,3)]
    cvd ← status %in% c("dead - heart or vascular",
                        "dead - cerebrovascular")
    rxn = as.numeric(rx) })
# Use transcan to compute optimal pre-transformations
ptrans ←    # See Figure 8.3
  transcan(~ sz + sg + ap + sbp + dbp +
           age + wt + hg + ekg + pf + bm + hx + dtime + rx,
           imputed=TRUE, transformed=TRUE,
           data=prostate, pl=FALSE, pr=FALSE)
# Use transcan single imputations
imp ← impute(ptrans, data=prostate, list.out=TRUE)
```

```
Imputed missing values with the following frequencies
 and stored them in variables with their original names:

 sz  sg age  wt ekg
  5  11   1   2   8
```

```
NAvars ← all.vars(~ sz + sg + age + wt + ekg)
for(x in NAvars) prostate[[x]] ← imp[[x]]
subset ← prostate$status %in% c("dead - heart or vascular",
   "dead - cerebrovascular","dead - prostatic ca")
trans ← ptrans$transformed[subset,]
psub  ← prostate[subset,]
```

11.4 Regression on Original Variables, Principal Components and Pretransformations

We first examine the performance of data reduction in predicting the cause of death, similar to what we did for survival time in Section 8.6. The first analyses assess how well PCs (on raw and transformed variables) predict the cause of death.

There are 127 cvds. We use the 15:1 rule of thumb discussed on P. 72 to justify using the first 8 PCs. ap is log-transformed because of its extreme distribution.

```
# Function to compute the first k PCs
ipc ← function(x, k=1, ...)
  princomp(x, ..., cor=TRUE)$scores[,1:k]
# Compute the first 8 PCs on raw variables then on
# transformed ones
pc8 ← ipc(~ sz + sg + log(ap) + sbp + dbp + age +
           wt + hg + ekg.norm + pfn + bm + hx + rxn + dtime,
           data=psub, k=8)
f8   ← lrm(cvd ~ pc8, data=psub)
pc8t ← ipc(trans, k=8)
f8t  ← lrm(cvd ~ pc8t, data=psub)
# Fit binary logistic model on original variables
f ← lrm(cvd ~ sz + sg + log(ap) + sbp + dbp + age +
         wt + hg + ekg + pf + bm + hx + rx + dtime, data=psub)
# Expand continuous variables using splines
g ← lrm(cvd ~ rcs(sz,4) + rcs(sg,4) + rcs(log(ap),4) +
         rcs(sbp,4) + rcs(dbp,4) + rcs(age,4) + rcs(wt,4) +
         rcs(hg,4) + ekg + pf + bm + hx + rx + rcs(dtime,4),
         data=psub)
# Fit binary logistic model on individual transformed var.
h ← lrm(cvd ~ trans, data=psub)
```

The five approaches to modeling the outcome are compared using AIC (where smaller is better).

```
c(f8=AIC(f8), f8t=AIC(f8t), f=AIC(f), g=AIC(g), h=AIC(h))
```

```
      f8      f8t       f       g       h
257.6573 254.5172 255.8545 263.8413 254.5317
```

Based on AIC, the more traditional model fitted to the raw data and assuming linearity for all the continuous predictors has only a slight chance of producing worse cross-validated predictive accuracy than other methods.

The chances are also good that effect estimates from this simple model will have competitive mean squared errors.

11.5 Description of Fitted Model

Here we describe the simple all-linear full model. Summary statistics and a Wald-ANOVA table are below, followed by partial effects plots with pointwise confidence bands, and odds ratios over default ranges of predictors.

```
print(f, latex=TRUE)
```

Logistic Regression Model

```
lrm(formula = cvd ~ sz + sg + log(ap) + sbp + dbp + age + wt +
    hg + ekg + pf + bm + hx + rx + dtime, data = psub)
```

		Model Likelihood Ratio Test		Discrimination Indexes		Rank Discrim. Indexes	
Obs	257	LR χ^2	144.39	R^2	0.573	C	0.893
FALSE	130	d.f.	21	g	2.688	D_{xy}	0.786
TRUE	127	$\Pr(>\chi^2) < 0.0001$		g_r	14.701	γ	0.787
$\max \lvert \frac{\partial \log L}{\partial \beta} \rvert$	6×10^{-11}			g_p	0.394	τ_a	0.395
				Brier	0.133		

	Coef	S.E.	Wald Z	$\Pr(>\lvert Z \rvert)$
Intercept	-4.5130	3.2210	-1.40	0.1612
sz	-0.0640	0.0168	-3.80	0.0001
sg	-0.2967	0.1149	-2.58	0.0098
ap	-0.3927	0.1411	-2.78	0.0054
sbp	-0.0572	0.0890	-0.64	0.5201
dbp	0.3917	0.1629	2.40	0.0162
age	0.0926	0.0286	3.23	0.0012
wt	-0.0177	0.0140	-1.26	0.2069
hg	0.0860	0.0925	0.93	0.3524
ekg=bngn	1.0781	0.8793	1.23	0.2202
ekg=rd&ec	-0.1929	0.6318	-0.31	0.7601
ekg=hbocd	-1.3679	0.8279	-1.65	0.0985
ekg=hrts	0.4365	0.4582	0.95	0.3407
ekg=MI	0.3039	0.5618	0.54	0.5886
pf=in bed < 50% daytime	0.9604	0.6956	1.38	0.1673
pf=in bed > 50% daytime	-2.3232	1.2464	-1.86	0.0623
bm	0.1456	0.5067	0.29	0.7738
hx	1.0913	0.3782	2.89	0.0039
rx=0.2 mg estrogen	-0.3022	0.4908	-0.62	0.5381
rx=1.0 mg estrogen	0.7526	0.5272	1.43	0.1534
rx=5.0 mg estrogen	0.6868	0.5043	1.36	0.1733
dtime	-0.0136	0.0107	-1.27	0.2040

```
an ← anova(f)
latex(an, file='', table.env=FALSE)
```

	χ^2	d.f.	P
sz	14.42	1	0.0001
sg	6.67	1	0.0098
ap	7.74	1	0.0054
sbp	0.41	1	0.5201
dbp	5.78	1	0.0162
age	10.45	1	0.0012
wt	1.59	1	0.2069
hg	0.86	1	0.3524
ekg	6.76	5	0.2391
pf	5.52	2	0.0632
bm	0.08	1	0.7738
hx	8.33	1	0.0039
rx	5.72	3	0.1260
dtime	1.61	1	0.2040
TOTAL	66.87	21	< 0.0001

```
plot(an)   # Figure 11.1
s ← f$stats
gamma.hat ← (s['Model L.R.'] - s['d.f.'])/s['Model L.R.']
```

```
dd ← datadist(psub); options(datadist='dd')
ggplot(Predict(f), sepdiscrete='vertical', vnames='names',
       rdata=psub,
       histSpike.opts=list(frac=function(f) .1*f/max(f) ))
       # Figure 11.2
```

```
plot(summary(f), log=TRUE)   # Figure 11.3
```

The van Houwelingen–Le Cessie heuristic shrinkage estimate (Equation 4.3) is $\hat{\gamma} = 0.85$, indicating that this model will validate on new data about 15% worse than on this dataset.

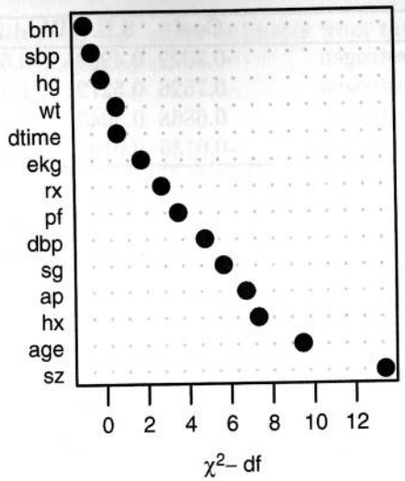


Fig. 11.1 Ranking of apparent importance of predictors of cause of death

11.6 Backwards Step-Down

Now use fast backward step-down (with total residual AIC as the stopping rule) to identify the variables that explain the bulk of the cause of death. Later validation will take this screening of variables into account.The greatly reduced model results in a simple nomogram.

```
fastbw(f)
```

```
 Deleted Chi-Sq d.f. P      Residual d.f. P      AIC
 ekg      6.76   5    0.2391   6.76     5   0.2391  -3.24
 bm       0.09   1    0.7639   6.85     6   0.3349  -5.15
 hg       0.38   1    0.5378   7.23     7   0.4053  -6.77
 sbp      0.48   1    0.4881   7.71     8   0.4622  -8.29
 wt       1.11   1    0.2932   8.82     9   0.4544  -9.18
 dtime    1.47   1    0.2253  10.29    10   0.4158  -9.71
 rx       5.65   3    0.1302  15.93    13   0.2528 -10.07
 pf       4.78   2    0.0915  20.71    15   0.1462  -9.29
 sg       4.28   1    0.0385  25.00    16   0.0698  -7.00
 dbp      5.84   1    0.0157  30.83    17   0.0209  -3.17

Approximate Estimates after Deleting Factors

              Coef    S.E. Wald Z         P
Intercept -3.74986 1.82887 -2.050 0.0403286
sz        -0.04862 0.01532 -3.174 0.0015013
ap        -0.40694 0.11117 -3.660 0.0002518
age        0.06000 0.02562  2.342 0.0191701
hx         0.86969 0.34339  2.533 0.0113198

Factors in Final Model

[1] sz  ap  age hx
```

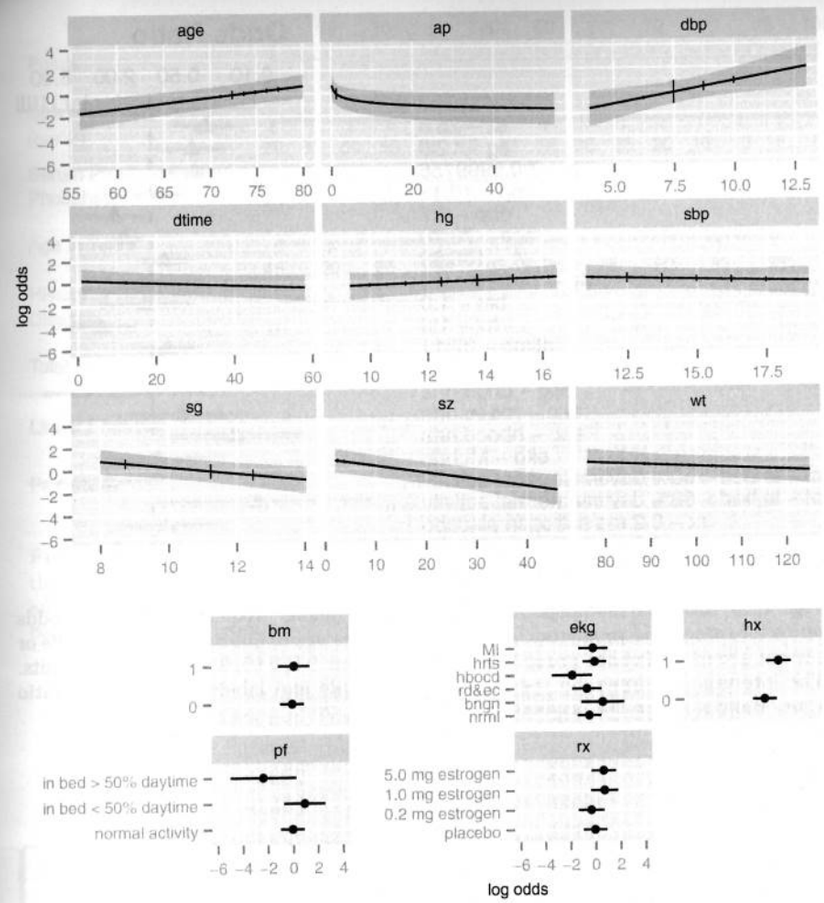



Fig. 11.2 Partial effects (log odds scale) in full model for cause of death, along with vertical line segments showing the raw data distribution of predictors

```
fred ← lrm(cvd ~ sz + log(ap) + age + hx, data=psub)
latex(fred, file='')
```

$$\text{Prob}\{\text{cvd}\} = \frac{1}{1+\exp(-X\beta)}, \quad \text{where}$$

$$X\hat{\beta} =$$
$$-5.009276 - 0.05510121\,\text{sz} - 0.509185\log(\text{ap}) + 0.0788052\,\text{age} + 1.070601\,\text{hx}$$

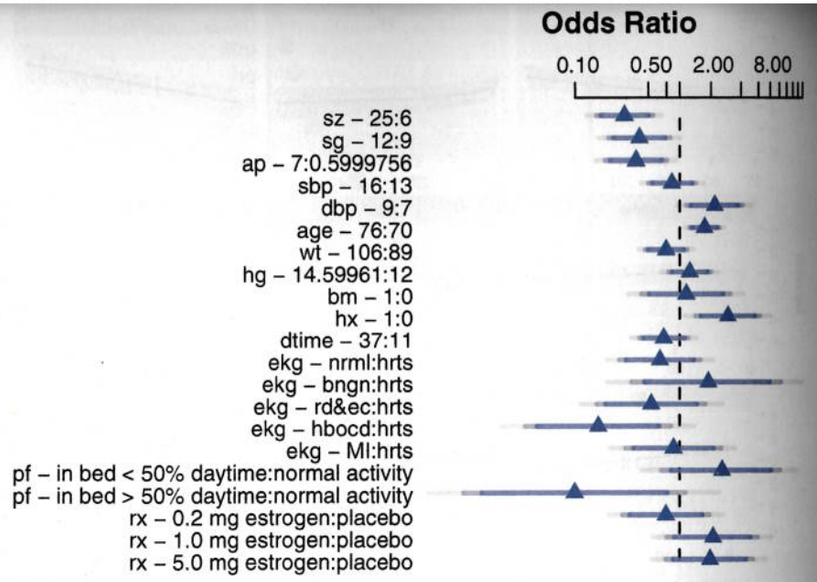


Fig. 11.3 Interquartile-range odds ratios for continuous predictors and simple odds ratios for categorical predictors. Numbers at left are upper quartile : lower quartile or current group : reference group. The bars represent 0.9, 0.95, 0.99 confidence limits. The intervals are drawn on the log odds ratio scale and labeled on the odds ratio scale. Ranges are on the original scale.

```
nom ← nomogram(fred, ap=c(.1, .5, 1, 5, 10, 50),
               fun=plogis, funlabel="Probability",
               fun.at=c(.01,.05,.1,.25,.5,.75,.9,.95,.99))
plot(nom, xfrac=.45)    # Figure 11.4
```

It is readily seen from this model that patients with a history of heart disease, and patients with less extensive prostate cancer are those more likely to die from `cvd` rather than from cancer. But beware that it is easy to overinterpret findings when using unpenalized estimation, and confidence intervals are too narrow. Let us use the bootstrap to study the uncertainty in the selection of variables and to penalize for this uncertainty when estimating predictive performance of the model. The variables selected in the first 20 bootstrap resamples are shown, making it obvious that the set of "significant" variables, i.e., the final model, is somewhat arbitrary.

```
f ← update(f, x=TRUE, y=TRUE)
v ← validate(f, B=200, bw=TRUE)
```

```
latex(v, B=20, digits=3)
```

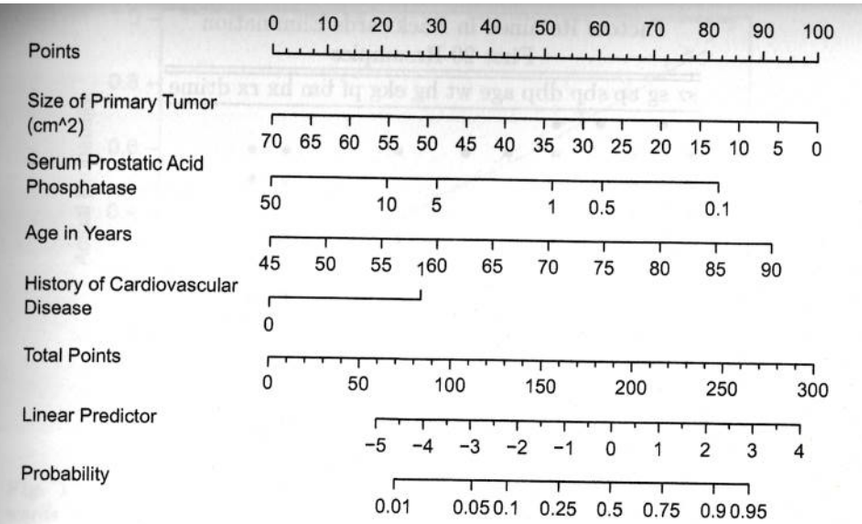
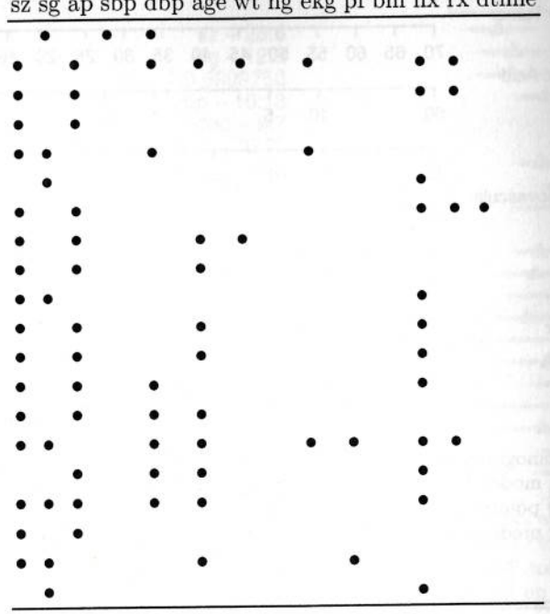


Fig. 11.4 Nomogram calculating $X\hat{\beta}$ and $\hat{P}$ for `cvd` as the cause of death, using the step-down model. For each predictor, read the points assigned on the 0–100 scale and add these points. Read the result on the `Total Points` scale and then read the corresponding predictions below it.

Index	Original Sample	Training Sample	Test Sample	Optimism	Corrected Index	n
D_{xy}	0.682	0.713	0.643	0.071	0.611	200
R^2	0.439	0.481	0.393	0.088	0.351	200
Intercept	0.000	0.000	−0.006	0.006	−0.006	200
Slope	1.000	1.000	0.811	0.189	0.811	200
$E_{\max}$	0.000	0.000	0.048	0.048	0.048	200
D	0.395	0.449	0.346	0.102	0.293	200
U	−0.008	−0.008	0.018	−0.026	0.018	200
Q	0.403	0.456	0.329	0.128	0.275	200
B	0.162	0.151	0.174	−0.022	0.184	200
g	1.932	2.213	1.756	0.457	1.475	200
g_p	0.341	0.355	0.320	0.035	0.306	200

Factors Retained in Backwards Elimination
First 20 Resamples



Frequencies of Numbers of Factors Retained

1	2	3	4	5	6	7	8	9	11	12
6	39	47	61	19	10	8	4	2	3	1

The slope shrinkage ($\hat{\gamma}$) is a bit lower than was estimated above. There is drop-off in all indexes. The estimated likely future predictive discrimination of the model as measured by Somers' D_{xy} fell from 0.682 to 0.611. The latter estimate is the one that should be claimed when describing model performance.

A nearly unbiased estimate of future calibration of the stepwise-derived model is given below.

```
cal ← calibrate(f, B=200, bw=TRUE)
plot(cal)    # Figure 11.5
```

The amount of overfitting seen in Figure 11.5 is consistent with the indexes produced by the `validate` function.

For comparison, consider a bootstrap validation of the full model without using variable selection.

```
vfull ← validate(f, B=200)
latex(vfull, digits=3)
```

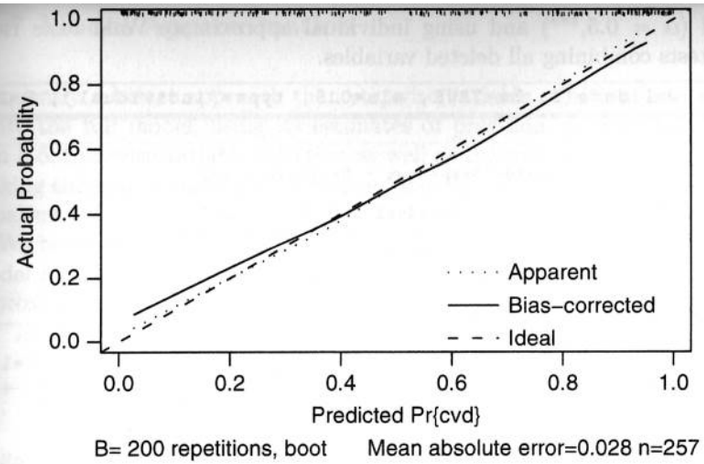


Fig. 11.5 Bootstrap overfitting–corrected calibration curve estimate for the backwards step-down cause of death logistic model, along with a rug plot showing the distribution of predicted risks. The smooth nonparametric calibration estimator (`loess`) is used.

Index	Original Sample	Training Sample	Test Sample	Optimism	Corrected Index	n
D_{xy}	0.786	0.833	0.738	0.095	0.691	200
R^2	0.573	0.641	0.501	0.140	0.433	200
Intercept	0.000	0.000	−0.013	0.013	−0.013	200
Slope	1.000	1.000	0.690	0.310	0.690	200
$E_{\max}$	0.000	0.000	0.085	0.085	0.085	200
D	0.558	0.653	0.468	0.185	0.373	200
U	−0.008	−0.008	0.051	−0.058	0.051	200
Q	0.566	0.661	0.417	0.244	0.322	200
B	0.133	0.115	0.150	−0.035	0.168	200
g	2.688	3.464	2.355	1.108	1.579	200
g_p	0.394	0.416	0.366	0.050	0.344	200

Compared to the validation of the full model, the step-down model has less optimism, but it started with a smaller D_{xy} due to loss of information from removing moderately important variables. The improvement in optimism was not enough to offset the effect of eliminating variables. If shrinkage were used with the full model, it would have better calibration and discrimination than the reduced model, since shrinkage does not diminish D_{xy}. Thus stepwise variable selection failed at delivering excellent predictive discrimination.

Finally, compare previous results with a bootstrap validation of a step-down model using a better significance level for a variable to stay in the

model ($\alpha = 0.5$, "bsy") and using individual approximate Wald tests rather than tests combining all deleted variables.

```
v5 ← validate(f, bw=TRUE, sls=0.5, type='individual', B=200)
```

```
                Backwards Step-down - Original Model

 Deleted Chi-Sq d.f. P      Residual d.f. P      AIC
 ekg     6.76    5   0.2391  6.76     5   0.2391  -3.24
 bm      0.09    1   0.7639  6.85     6   0.3349  -5.15
 hg      0.38    1   0.5378  7.23     7   0.4053  -6.77
 sbp     0.48    1   0.4881  7.71     8   0.4622  -8.29
 wt      1.11    1   0.2932  8.82     9   0.4544  -9.18
 dtime   1.47    1   0.2253 10.29    10   0.4158  -9.71
 rx      5.65    3   0.1302 15.93    13   0.2528 -10.07

Approximate Estimates after Deleting Factors

                                Coef    S.E. Wald Z        P
Intercept                    -4.86308 2.67292 -1.819 0.068852
sz                           -0.05063 0.01581 -3.202 0.001366
sg                           -0.28038 0.11014 -2.546 0.010903
ap                           -0.24838 0.12369 -2.008 0.044629
dbp                           0.28288 0.13036  2.170 0.030008
age                           0.08502 0.02690  3.161 0.001572
pf=in bed < 50% daytime       0.81151 0.66376  1.223 0.221485
pf=in bed > 50% daytime      -2.19885 1.21212 -1.814 0.069670
hx                            0.87834 0.35203  2.495 0.012592

Factors in Final Model

[1] sz  sg  ap  dbp age pf  hx
```

```
latex(v5, digits=3, B=0)
```

Index	Original Sample	Training Sample	Test Sample	Optimism	Corrected Index	n
D_{xy}	0.739	0.801	0.716	0.085	0.654	200
R^2	0.517	0.598	0.481	0.117	0.400	200
Intercept	0.000	0.000	−0.008	0.008	−0.008	200
Slope	1.000	1.000	0.745	0.255	0.745	200
$E_{\max}$	0.000	0.000	0.067	0.067	0.067	200
D	0.486	0.593	0.444	0.149	0.337	200
U	−0.008	−0.008	0.033	−0.040	0.033	200
Q	0.494	0.601	0.411	0.190	0.304	200
B	0.147	0.125	0.156	−0.030	0.177	200
g	2.351	2.958	2.175	0.784	1.567	200
g_p	0.372	0.401	0.358	0.043	0.330	200

The performance statistics are midway between the full model and the smaller stepwise model.

11.7 Model Approximation

Frequently a better approach than stepwise variable selection is to approximate the full model, using its estimates of precision, as discussed in Section 5.5. Stepwise variable selection as well as regression trees are useful for making the approximations, and the sacrifice in predictive accuracy is always apparent.

We begin by computing the "gold standard" linear predictor from the full model fit ($R^2 = 1.0$), then running backwards step-down OLS regression to approximate it.

```
lp ← predict(f)    # Compute linear predictor from full model
# Insert sigma=1 as otherwise sigma=0 will cause problems
a ← ols(lp ~ sz + sg + log(ap) + sbp + dbp + age + wt +
          hg + ekg + pf + bm + hx + rx + dtime, sigma=1,
          data=psub)
# Specify silly stopping criterion to remove all variables
s ← fastbw(a, aics=10000)
betas ← s$Coefficients    # matrix, rows=iterations
X     ← cbind(1, f$x)     # design matrix
# Compute the series of approximations to lp
ap ← X %*% t(betas)
# For each approx. compute approximation R^2 and ratio of
# likelihood ratio chi-square for approximate model to that
# of original model
m ← ncol(ap) - 1    # all but intercept-only model
r2 ← frac ← numeric(m)
fullchisq ← f$stats['Model L.R.']
for(i in 1:m) {
  lpa ← ap[,i]
  r2[i] ← cor(lpa, lp)^2
  fapprox ← lrm(cvd ~ lpa, data=psub)
  frac[i] ← fapprox$stats['Model L.R.'] / fullchisq
  }    # Figure 11.6:
plot(r2, frac, type='b',
     xlab=expression(paste('Approximation ', R^2)),
     ylab=expression(paste('Fraction of ',
          chi^2, ' Preserved')))
abline(h=.95, col=gray(.83)); abline(v=.95, col=gray(.83))
abline(a=0, b=1, col=gray(.83))
```

After 6 deletions, slightly more than 0.05 of both the LR χ^2 and the approximation R^2 are lost (see Figure 11.6). Therefore we take as our approximate model the one that removed 6 predictors. The equation for this model is below, and its nomogram is in Figure 11.7.

```
fapprox ← ols(lp ~ sz + sg + log(ap) + age + ekg + pf + hx +
              rx, data=psub)
fapprox$stats['R2']    # as a check
```

R2 0.9453396

```
latex(fapprox, file='')
```

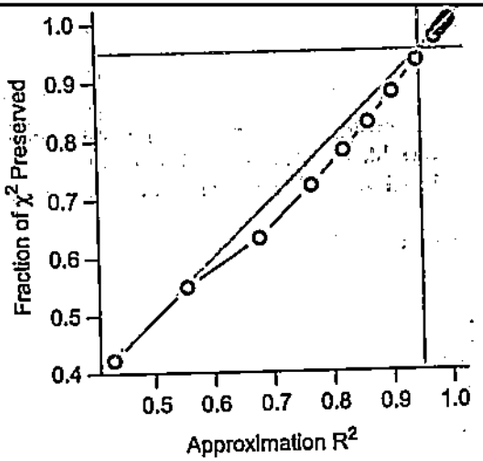


Fig. 11.6 Fraction of explainable variation (full model LR χ^2) in cvd that was explained by approximate models, along with approximation accuracy (x-axis)

$$\mathrm{E}(\mathrm{lp}) = X\beta, \quad \text{where}$$

$X\hat{\beta} =$
$-2.868303 - 0.06233241\,\mathrm{sz} - 0.3157901\,\mathrm{sg} - 0.3834479\,\log(\mathrm{ap}) + 0.09089393\,\mathrm{age}$
$+1.396922[\mathrm{bngn}] + 0.06275034[\mathrm{rd\&ec}] - 1.24892[\mathrm{hbocd}] + 0.6511938[\mathrm{hrts}]$
$+0.3236771[\mathrm{MI}]$
$+1.116028[\text{in bed} < 50\% \text{ daytime}] - 2.436734[\text{in bed} > 50\% \text{ daytime}]$
$+1.05316\,\mathrm{hx}$
$-0.3888534[0.2 \text{ mg estrogen}] + 0.6920495[1.0 \text{ mg estrogen}]$
$+0.7834498[5.0 \text{ mg estrogen}]$

and $[c] = 1$ if subject is in group c, 0 otherwise.

```
nom ← nomogram(fapprox, ap=c(.1, .5, 1, 5, 10, 20, 30, 40),
               fun=plogis, funlabel="Probability",
               lp.at=(-5):4,
          fun.lp.at=qlogis(c(.01,.05,.25,.5,.75,.96,.99)))
plot(nom, xfrac=.45)    # Figure 11.7
```

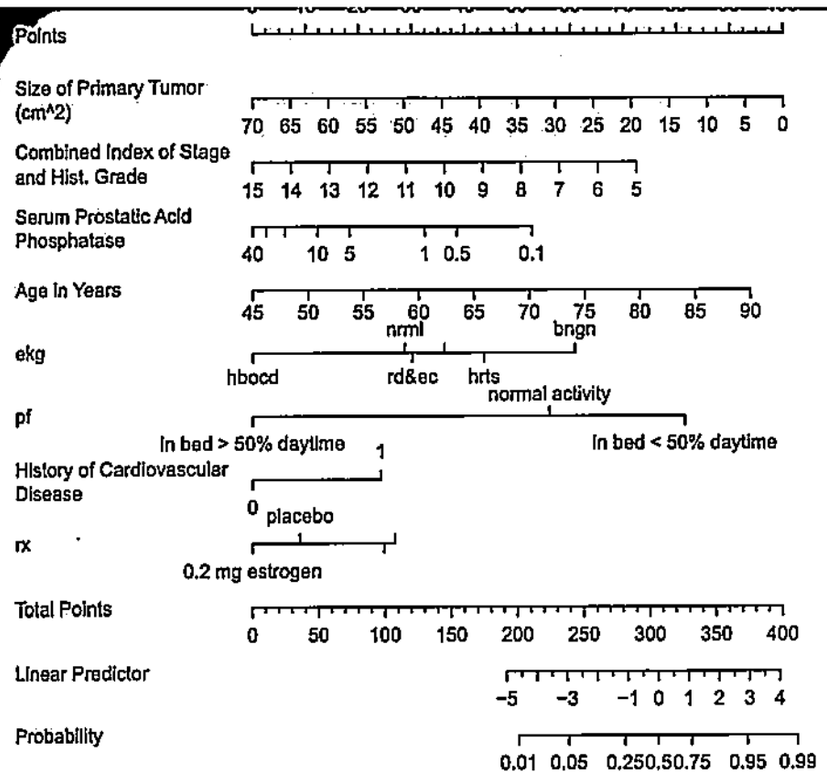


Fig. 11.7 Nomogram for predicting the probability of cvd based on the approximate model