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

[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.

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 ones that are most different from one another. So if using partial tests or rank correlation to reduce degrees of freedom can harm the model, one might argue that it is fair to allow this strategy to also benefit the analysis.

When collinearities or confounding are not problematic, a quicker approach based on pairwise measures of association can be useful. This approach will not have numerical problems (e.g., singular covariance matrix). When Y is binary or continuous (but not censored), a good general-purpose measure of association that is useful in making decisions about the number of parameters to devote to a predictor is an extension of Spearman's ρ rank correlation. This is the ordinary R^2 from predicting the rank of Y based on the rank of X and the square of the rank of X. This ρ^2 will detect not only nonlinear relationships (as will ordinary Spearman ρ) but some non-monotonic ones as well. [2] It is important that the ordinary Spearman ρ not be computed, as this would tempt the analyst to simplify the regression function (towards monotonicity) if the generalized ρ^2 does not significantly exceed the square of the ordinary Spearman ρ. For categorical predictors, ranks are not squared but instead the predictor is represented by a series of dummy variables. The resulting ρ^2 is related to the Kruskal–Wallis test. See p. 460 for an example. Note that bivariable correlations can be misleading if marginal relationships vary greatly from ones obtained after adjusting for other predictors. [3]

Once one expands a predictor into linear and nonlinear terms and estimates the coefficients, the best way to understand the relationship between predictors and response is to graph this estimated relationship[c]. If the plot appears almost linear or the test of nonlinearity is very insignificant there is a temptation to simplify the model. The Grambsch and O'Brien result described in Section 2.6 demonstrates why this is a bad idea.

From the above discussion a general principle emerges. Whenever the response variable is informally or formally linked, in an unmasked fashion, to particular parameters that may be deleted from the model, special adjustments must be made in P-values, standard errors, test statistics, and confidence limits, in order for these statistics to have the correct interpretation. Examples of strategies that are improper without special adjustments (e.g., using the bootstrap) include examining a frequency table or scatterplot to decide that an association is too weak for the predictor to be included in the model at all or to decide that the relationship appears so linear that all nonlinear terms should be omitted. It is also valuable to consider the reverse situation; that is, one posits a simple model and then additional analysis or outside subject matter information makes the analyst want to generalize the model. Once the model is generalized (e.g., nonlinear terms are added), the test of association can be recomputed using multiple d.f. So another general principle is that when one makes the model more complex, the d.f. properly increases and the new test statistics for association have the claimed

[c] One can also perform a joint test of all parameters associated with nonlinear effects. This can be useful in demonstrating to the reader that some complexity was actually needed.

distribution. Thus moving from simple to more complex models presents no problems other than conservatism if the new complex components are truly unnecessary.

4.2 Checking Assumptions of Multiple Predictors Simultaneously

Before developing a multivariable model one must decide whether the assumptions of each continuous predictor can be verified by ignoring the effects of all other potential predictors. In some cases, the shape of the relationship between a predictor and the property of response will be different if an adjustment is made for other correlated factors when deriving regression estimates. Also, failure to adjust for an important factor can frequently alter the nature of the distribution of Y. Occasionally, however, it is unwieldy to deal simultaneously with all predictors at each stage in the analysis, and instead the regression function shapes are assessed separately for each continuous predictor.

4.3 Variable Selection

The material covered to this point dealt with a prespecified list of variables to be included in the regression model. For reasons of developing a concise model or because of a fear of collinearity or of a false belief that it is not legitimate to include "insignificant" regression coefficients when presenting results to the intended audience, stepwise variable selection is very commonly employed. Variable selection is used when the analyst is faced with a series of potential predictors but does not have (or use) the necessary subject matter knowledge to enable her to prespecify the "important" variables to include in the model. But using Y to compute P-values to decide which variables to include is similar to using Y to decide how to pool treatments in a five–treatment randomized trial, and then testing for global treatment differences using fewer than four degrees of freedom.

Stepwise variable selection has been a very popular technique for many years, but if this procedure had just been proposed as a statistical method, it would most likely be rejected because it violates every principle of statistical estimation and hypothesis testing. Here is a summary of the problems with this method.

1. It yields R^2 values that are biased high.
2. The ordinary F and χ^2 test statistics do not have the claimed distribution[d].[234] Variable selection is based on methods (e.g., F tests for nested models) that were intended to be used to test only prespecified hypotheses.
3. The method yields standard errors of regression coefficient estimates that are biased low and confidence intervals for effects and predicted values that are falsely narrow.[16]
4. It yields P-values that are too small (i.e., there are severe multiple comparison problems) and that do not have the proper meaning, and the proper correction for them is a very difficult problem.
5. It provides regression coefficients that are biased high in absolute value and need shrinkage. Even if only a single predictor were being analyzed and one only reported the regression coefficient for that predictor if its association with Y were "statistically significant," the estimate of the regression coefficient $\hat{\beta}$ is biased (too large in absolute value). To put this in symbols for the case where we obtain a positive association ($\hat{\beta} > 0$), $E(\hat{\beta}|P < 0.05, \hat{\beta} > 0) > \beta$.[100]
6. In observational studies, variable selection to determine confounders for adjustment results in residual confounding[241].
7. Rather than solving problems caused by collinearity, variable selection is made arbitrary by collinearity.
8. It allows us to not think about the problem.

The problems of P-value-based variable selection are exacerbated when the analyst (as she so often does) interprets the final model as if it were prespecified. Copas and Long[125] stated one of the most serious problems with stepwise modeling eloquently when they said, "The choice of the variables to be included depends on estimated regression coefficients rather than their true values, and so X_j is more likely to be included if its regression coefficient is over-estimated than if its regression coefficient is underestimated." Derksen and Keselman[155] studied stepwise variable selection, backward elimination, and forward selection, with these conclusions:

> 1. "The degree of correlation between the predictor variables affected the frequency with which authentic predictor variables found their way into the final model.
> 2. The number of candidate predictor variables affected the number of noise variables that gained entry to the model.
> 3. The size of the sample was of little practical importance in determining the number of authentic variables contained in the final model.

[d] Lockhart et al.[425] provide an example with $n = 100$ and 10 orthogonal predictors where all true βs are zero. The test statistic for the first variable to enter has type I error of 0.39 when the nominal α is set to 0.05, in line with what one would expect with multiple testing using $1 - 0.95^{10} = 0.40$.

4. The population multiple coefficient of determination could be faithfully estimated by adopting a statistic that is adjusted by the total number of candidate predictor variables rather than the number of variables in the final model."

They found that variables selected for the final model represented noise 0.20 to 0.74 of the time and that the final model usually contained less than half of the actual number of authentic predictors. Hence there are many reasons for using methods such as full-model fits or data reduction, instead of using any stepwise variable selection algorithm.

If stepwise selection must be used, a global test of no regression should be made before proceeding, simultaneously testing all candidate predictors and having degrees of freedom equal to the number of candidate variables (plus any nonlinear or interaction terms). If this global test is not significant, selection of individually significant predictors is usually not warranted.

The method generally used for such variable selection is forward selection of the most significant candidate or backward elimination of the least significant predictor in the model. One of the recommended stopping rules is based on the "residual χ^2" with degrees of freedom equal to the number of candidate variables remaining at the current step. The residual χ^2 can be tested for significance (if one is able to forget that because of variable selection this statistic does not have a χ^2 distribution), or the stopping rule can be based on Akaike's information criterion (AIC[33]), here residual $\chi^2 - 2\times$ d.f.[257] Of course, use of more insight from knowledge of the subject matter will generally improve the modeling process substantially. It must be remembered that no currently available stopping rule was developed for data-driven variable selection. Stopping rules such as AIC or Mallows' C_p are intended for comparing a limited number of *prespecified* models [66, Section 1.3][347][e]. 4

If the analyst insists on basing the stopping rule on P-values, the optimum (in terms of predictive accuracy) α to use in deciding which variables to include in the model is $\alpha = 1.0$ unless there are a few powerful variables and several completely irrelevant variables. A reasonable α that does allow for deletion of *some* variables is $\alpha = 0.5$.[589] These values are far from the traditional choices of $\alpha = 0.05$ or 0.10. 5

[e] AIC works successfully when the models being entertained are on a progression defined by a single parameter, e.g. a common shrinkage coefficient or the single number of knots to be used by *all* continuous predictors. AIC can also work when the model that is best by AIC is much better than the runner-up so that if the process were bootstrapped the same model would almost always be found. When used for one variable at a time variable selection. AIC is just a restatement of the P-value, and as such, doesn't solve the severe problems with stepwise variable selection other than forcing us to use slightly more sensible α values. Burnham and Anderson[84] recommend selection based on AIC for a limited number of theoretically well-founded models. Some statisticians try to deal with multiplicity problems caused by stepwise variable selection by making α smaller than 0.05. This increases bias by giving variables whose effects are estimated with error a greater relative chance of being selected. Variable selection does not compete well with shrinkage methods that simultaneously model all potential predictors.

6 Even though forward stepwise variable selection is the most commonly used method, the step-down method is preferred for the following reasons.

1. It usually performs better than forward stepwise methods, especially when collinearity is present.[437]
2. It makes one examine a full model fit, which is the only fit providing accurate standard errors, error mean square, and P-values.
3. The method of Lawless and Singhal[385] allows extremely efficient step-down modeling using Wald statistics, in the context of any fit from least squares or maximum likelihood. This method requires passing through the data matrix only to get the initial full fit.

For a given dataset, bootstrapping (Efron et al.[150, 172, 177, 178]) can help decide between using full and reduced models. Bootstrapping can be done on the whole model and compared with bootstrapped estimates of predictive accuracy based on stepwise variable selection for each resample. Unless most predictors are either very significant or clearly unimportant, the full model usually outperforms the reduced model.

Full model fits have the advantage of providing meaningful confidence intervals using standard formulas. Altman and Andersen[16] gave an example in which the lengths of confidence intervals of predicted survival probabilities were 60% longer when bootstrapping was used to estimate the simultaneous effects of variability caused by variable selection and coefficient estimation, as compared with confidence intervals computed ignoring how a “final” model came to be. 7 On the other hand, models developed on full fits after data reduction will be optimum in many cases. 8

In some cases you may want to use the full model for prediction and variable selection for a “best bet” parsimonious list of independently important predictors. This could be accompanied by a list of variables selected in 50 bootstrap samples to demonstrate the imprecision in the “best bet.”

Sauerbrei and Schumacher[541] present a method to use bootstrapping to actually select the set of variables. However, there are a number of drawbacks to this approach[35]:

1. The choice of an α cutoff for determining whether a variable is retained in a given bootstrap sample is arbitrary.
2. The choice of a cutoff for the proportion of bootstrap samples for which a variable is retained, in order to include that variable in the final model, is somewhat arbitrary.
3. Selection from among a set of correlated predictors is arbitrary, and all highly correlated predictors may have a low bootstrap selection frequency. It may be the case that none of them will be selected for the final model even though when considered individually each of them may be highly significant.

4. By using the bootstrap to choose variables, one must use the double bootstrap to resample the entire modeling process in order to validate the model and to derive reliable confidence intervals. This may be computationally prohibitive.
5. The bootstrap did not improve upon traditional backward stepdown variable selection. Both methods fail at identifying the "correct" variables.

For some applications the list of variables selected may be stabilized by grouping variables according to subject matter considerations or empirical correlations and testing each related group with a multiple degree of freedom test. Then the entire group may be kept or deleted and, if desired, groups that are retained can be summarized into a single variable or the most accurately measured variable within the group can replace the group. See Section 4.7 for more on this.

Kass and Raftery[337] showed that Bayes factors have several advantages in variable selection, including the selection of less complex models that may agree better with subject matter knowledge. However, as in the case with more traditional stopping rules, the final model may still have regression coefficients that are too large. This problem is solved by Tibshirani's *lasso* method,[608, 609] which is a penalized estimation technique in which the estimated regression coefficients are constrained so that the sum of their scaled absolute values falls below some constant k chosen by cross-validation. This kind of constraint forces some regression coefficient estimates to be exactly zero, thus achieving variable selection while shrinking the remaining coefficients toward zero to reflect the overfitting caused by data-based model selection.

A final problem with variable selection is illustrated by comparing this approach with the sensible way many economists develop regression models. Economists frequently use the strategy of deleting only those variables that are "insignificant" and whose regression coefficients have a nonsensible direction. Standard variable selection on the other hand yields biologically implausible findings in many cases by setting certain regression coefficients exactly to zero. In a study of survival time for patients with heart failure, for example, it would be implausible that patients having a specific symptom live exactly as long as those without the symptom just because the symptom's regression coefficient was "insignificant." The lasso method shares this difficulty with ordinary variable selection methods and with any method that in the Bayesian context places nonzero prior probability on β being *exactly* zero. 9

Many papers claim that there were insufficient data to allow for multivariable modeling, so they did "univariable screening" wherein only "significant" variables (i.e., those that are separately significantly associated with Y) were entered into the model.[f] This is just a forward stepwise variable selection in

[f] This is akin to doing a t-test to compare the two treatments (out of 10, say) that are apparently most different from each other.

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 11 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.

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" 12 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"

[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]

where s^2 is the sample variance and $\chi^2_{\alpha,n-1}$ is the α critical value of the χ^2 distribution with $n-1$ degrees of freedom. We take the fold-change or multiplicative margin of error (MMOE) for estimating σ to be

$$\sqrt{\max(\frac{\chi^2_{1-\alpha/2,n-1}}{n-1}, \frac{n-1}{\chi^2_{\alpha/2,n-1}})} \tag{4.2}$$

To achieve a MMOE of no worse than 1.2 with 0.95 confidence when estimating σ requires a sample size of 70 subjects.

The linear model case is useful for examining $n:p$ ratio another way. As discussed in the next section, R^2_{adj} is a nearly unbiased estimate of R^2, i.e., is not inflated by overfitting if the value used for p is "honest", i.e., includes all variables screened. We can ask the question "for a given R^2, what ratio of $n:p$ is required so that R^2_{adj} does not drop by more than a certain relative or absolute amount from the value of R^2?" This assessment takes into account that higher signal:noise ratios allow fitting more variables. For example, with low R^2 a 100:1 ratio of $n:p$ may be required to prevent R^2 from dropping by more $\frac{1}{10}$ or by an absolute amount of 0.01. A 15:1 rule would prevent R^2 from dropping by more than 0.075 for low R^2 (Figure 4.1).

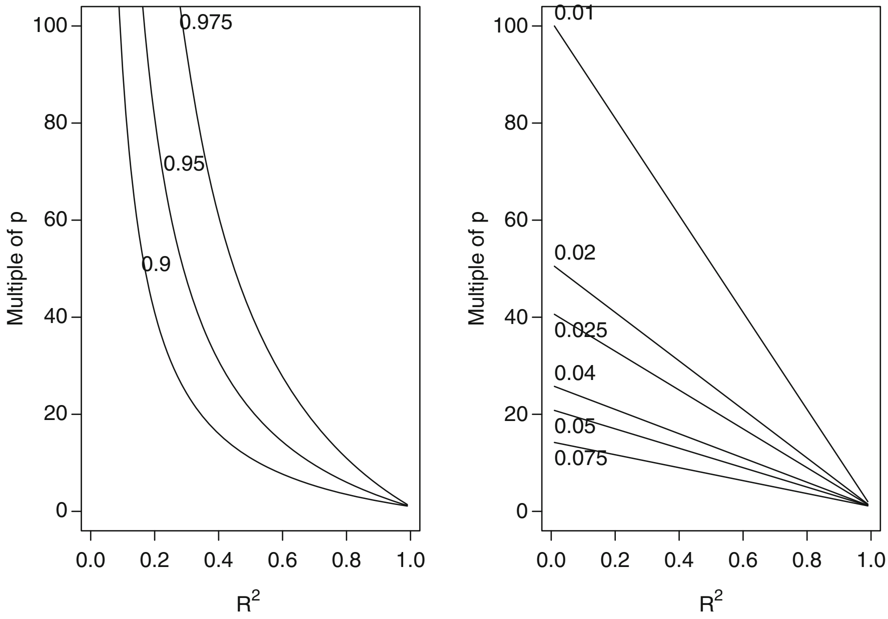

Fig. 4.1 Multiple of p that n must be to achieve a relative drop from R^2 to R^2_{adj} by the indicated relative factor (left panel, 3 factors) or absolute difference (right panel, 6 decrements)

4.5 Shrinkage

The term *shrinkage* is used in regression modeling to denote two ideas. The first meaning relates to the slope of a *calibration plot*, which is a plot of observed responses against predicted responses[k]. When a dataset is used to fit the model parameters as well as to obtain the calibration plot, the usual estimation process will force the slope of observed versus predicted values to be one. When, however, parameter estimates are derived from one dataset and then applied to predict outcomes on an independent dataset, overfitting will cause the slope of the calibration plot (i.e., the *shrinkage factor*) to be less than one, a result of *regression to the mean*. Typically, low predictions will be too low and high predictions too high. Predictions near the mean predicted value will usually be quite accurate. The second meaning of *shrinkage* is a statistical estimation method that preshrinks regression coefficients towards zero so that the calibration plot for new data will not need shrinkage as its calibration slope will be one.

We turn first to shrinkage as an adverse result of traditional modeling. In ordinary linear regression, we know that all of the coefficient estimates are exactly unbiased estimates of the true effect when the model fits. Isn't the existence of shrinkage and overfitting implying that there is some kind of bias in the parameter estimates? The answer is no because each separate coefficient has the desired expectation. The problem lies in how we use the coefficients. We tend not to pick out coefficients at random for interpretation but we tend to highlight very small and very large coefficients.

A simple example may suffice. Consider a clinical trial with 10 randomly assigned treatments such that the patient responses for each treatment are normally distributed. We can do an ANOVA by fitting a multiple regression model with an intercept and nine dummy variables. The intercept is an unbiased estimate of the mean response for patients on the first treatment, and each of the other coefficients is an unbiased estimate of the difference in mean response between the treatment in question and the first treatment. $\hat{\beta}_0 + \hat{\beta}_1$ is an unbiased estimate of the mean response for patients on the second treatment. But if we plotted the predicted mean response for patients against the observed responses from new data, the slope of this calibration plot would typically be smaller than one. This is because in making this plot we are not picking coefficients at random but we are sorting the coefficients into ascending order. The treatment group having the lowest sample mean response will usually have a higher mean in the future, and the treatment group having the highest sample mean response will typically have a lower mean in the future. The sample mean of the group having the highest sample mean is *not* an unbiased estimate of its population mean.

[k] An even more stringent assessment is obtained by stratifying calibration curves by predictor settings.

As an illustration, let us draw 20 samples of size $n = 50$ from a uniform distribution for which the true mean is 0.5. Figure 4.2 displays the 20 means sorted into ascending order, similar to plotting Y versus $\hat{Y} = X\hat{\beta}$ based on least squares after sorting by $X\hat{\beta}$. Bias in the very lowest and highest estimates is evident.

```
set.seed(123)
n ← 50
y ← runif(20*n)
group ← rep(1:20,each=n)
ybar ← tapply(y, group, mean)
ybar ← sort(ybar)
plot(1:20, ybar, type='n', axes=FALSE, ylim=c(.3,.7),
     xlab='Group', ylab='Group Mean')
lines(1:20, ybar)
points(1:20, ybar, pch=20, cex=.5)
axis(2)
axis(1, at=1:20, labels=FALSE)
for(j in 1:20) axis(1, at=j, labels=names(ybar)[j])
abline(h=.5, col=gray(.85))
```

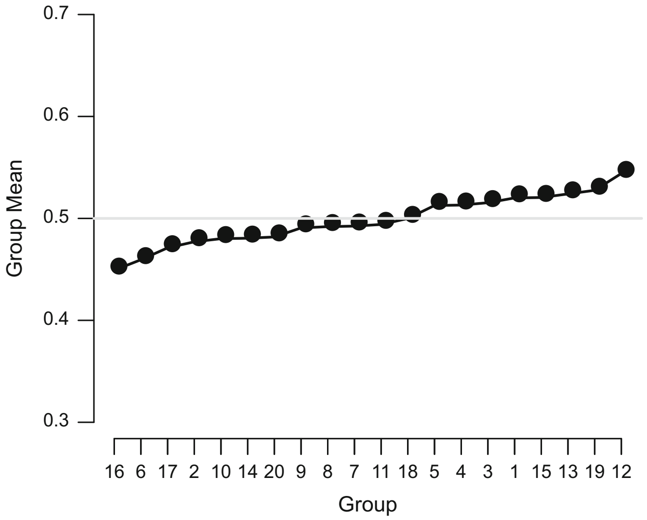

Fig. 4.2 Sorted means from 20 samples of size 50 from a uniform $[0, 1]$ distribution. The reference line at 0.5 depicts the true population value of all of the means.

When we want to highlight a treatment that is not chosen at random (or a priori), the data-based selection of that treatment needs to be compensated for in the estimation process.[1] It is well known that the use of shrinkage

[1] It is interesting that researchers are quite comfortable with adjusting P-values for post hoc selection of comparisons using, for example, the Bonferroni inequality, but they do not realize that post hoc selection of comparisons also biases point estimates.

methods such as the James–Stein estimator to pull treatment means toward the grand mean over all treatments results in estimates of treatment-specific means that are far superior to ordinary stratified means.[176]

Turning from a cell means model to the general case where predicted values are general linear combinations $X\hat{\beta}$, the slope γ of properly transformed responses Y against $X\hat{\beta}$ (sorted into ascending order) will be less than one on new data. Estimation of the shrinkage coefficient γ allows quantification of the amount of overfitting present, and it allows one to estimate the likelihood that the model will reliably predict new observations. van Houwelingen and le Cessie [633, Eq. 77] provided a heuristic shrinkage estimate that has worked well in several examples:

$$\hat{\gamma} = \frac{\text{model } \chi^2 - p}{\text{model } \chi^2}, \tag{4.3}$$

where p is the total degrees of freedom for the predictors and model χ^2 is 13
the likelihood ratio χ^2 statistic for testing the joint influence of all predictors simultaneously (see Section 9.3.1). For ordinary linear models, van Houwelingen and le Cessie proposed a shrinkage factor $\hat{\gamma}$ that can be shown to equal $\frac{n-p-1}{n-1}\frac{R^2_{\text{adj}}}{R^2}$, where the adjusted R^2 is given by 14

$$R^2_{\text{adj}} = 1 - (1 - R^2)\frac{n-1}{n-p-1}. \tag{4.4}$$

For such linear models with an intercept β_0, the shrunken estimate of β is

$$\begin{aligned} \hat{\beta}^s_0 &= (1-\hat{\gamma})\overline{Y} + \hat{\gamma}\hat{\beta}_0 \\ \hat{\beta}^s_j &= \hat{\gamma}\hat{\beta}_j, j = 1, \ldots, p, \end{aligned} \tag{4.5}$$

where $\overline{Y}$ is the mean of the response vector. Again, when stepwise fitting is used, the p in these equations is much closer to the number of *candidate* degrees of freedom rather than the number in the "final" model. See Section 5.3 15
for methods of estimating γ using the bootstrap (p. 115) or cross-validation.

Now turn to the second usage of the term *shrinkage*. Just as clothing is sometimes preshrunk so that it will not shrink further once it is purchased, better calibrated predictions result when shrinkage is built into the estimation process in the first place. The object of shrinking regression coefficient estimates is to obtain a shrinkage coefficient of $\gamma = 1$ on new data. Thus by somewhat discounting $\hat{\beta}$ we make the model underfitted on the data at hand (i.e., apparent $\gamma < 1$) so that on new data extremely low or high predictions are correct.

Ridge regression[388, 633] is one technique for placing restrictions on the parameter estimates that results in shrinkage. A *ridge parameter* must be chosen to control the amount of shrinkage. Penalized maximum likelihood estimation,[237, 272, 388, 639] a generalization of ridge regression, is a general shrinkage 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 16
as VIF are not very informative as some variables are algebraically connected to each other.

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

as principal surfaces, sliced inverse regression, variable clustering, or ordinary cluster analysis on a measure of similarity between variables. 18

See Chapters 8 and 14 for detailed case studies in data reduction.

4.7.1 Redundancy Analysis

There are many approaches to data reduction. One rigorous approach involves removing predictors that are easily predicted from other predictors, using flexible parametric additive regression models. This approach is unlikely to result in a major reduction in the number of regression coefficients to estimate against Y, but will usually provide insights useful for later data reduction over and above the insights given by methods based on pairwise correlations instead of multiple R^2.

The `Hmisc` `redun` function implements the following redundancy checking algorithm.

- Expand each continuous predictor into restricted cubic spline basis functions. Expand categorical predictors into dummy variables.
- Use OLS to predict each predictor with all component terms of all remaining predictors (similar to what the `Hmisc` `transcan` function does). When the predictor is expanded into multiple terms, use the first canonical variate[m].
- Remove the predictor that can be predicted from the remaining set with the highest adjusted or regular R^2.
- Predict all remaining predictors from their complement.
- Continue in like fashion until no variable still in the list of predictors can be predicted with an R^2 or adjusted R^2 greater than a specified threshold or until dropping the variable with the highest R^2 (adjusted or ordinary) would cause a variable that was dropped earlier to no longer be predicted at the threshold from the now smaller list of predictors.

Special consideration must be given to categorical predictors. One way to consider a categorical variable redundant is if a linear combination of dummy variables representing it can be predicted from a linear combination of other variables. For example, if there were 4 cities in the data and each city's rainfall was also present as a variable, with virtually the same rainfall reported for all observations for a city, city would be redundant given rainfall (or vice-versa). If two cities had the same rainfall, 'city' might be declared redundant even though tied cities might be deemed non-redundant in another setting. A second, more stringent way to check for redundancy of a categorical predictor is to ascertain whether all dummy variables created from the predictor are individually redundant. The `redun` function implements both approaches.

19 Examples of use of `redun` are given in two case studies.

[m] There is an option to force continuous variables to be linear when they are being predicted.

4.7.2 Variable Clustering

Although the use of subject matter knowledge is usually preferred, statistical clustering techniques can be useful in determining independent dimensions that are described by the entire list of candidate predictors. Once each dimension is scored (see below), the task of regression modeling is simplified, and one quits trying to separate the effects of factors that are measuring the same phenomenon. One type of variable clustering[539] is based on a type of oblique-rotation principal component (PC) analysis that attempts to separate variables so that the first PC of each group is representative of that group (the first PC is the linear combination of variables having maximum variance subject to normalization constraints on the coefficients[142, 144]). Another approach, that of doing a hierarchical cluster analysis on an appropriate similarity matrix (such as squared correlations) will often yield the same results. For either approach, it is often advisable to use robust (e.g., rank-based) measures for continuous variables if they are skewed, as skewed variables can greatly affect ordinary correlation coefficients. Pairwise deletion of missing values is also advisable for this procedure—casewise deletion can result in a small biased sample. 20

When variables are not monotonically related to each other, Pearson or Spearman squared correlations can miss important associations and thus are not always good similarity measures. A general and robust similarity measure is Hoeffding's D,[295] which for two variables X and Y is a measure of the agreement between $F(x, y)$ and $G(x)H(y)$, where G, H are marginal cumulative distribution functions and F is the joint CDF. The D statistic will detect a wide variety of dependencies between two variables.

See pp. 330 and 458 for examples of variable clustering. 21

4.7.3 Transformation and Scaling Variables Without Using Y

Scaling techniques often allow the analyst to reduce the number of parameters to fit by estimating transformations for each predictor using only information about associations with other predictors. It may be advisable to cluster variables before scaling so that patterns are derived only from variables that are related. For purely categorical predictors, methods such as correspondence analysis (see, for example, [108,139,239,391,456]) can be useful for data reduction. Often one can use these techniques to scale multiple dummy variables into a few dimensions. For mixtures of categorical and continuous predictors, qualitative principal component analysis such as the *maximum total variance* (MTV) method of Young et al.[456, 680] is useful. For the special case of representing a series of variables with one PC, the MTV method is quite easy to implement.

1. Compute PC_1, the first PC of the variables to reduce $X_1, \ldots, X_q$ using the correlation matrix of Xs.
2. Use ordinary linear regression to predict PC_1 on the basis of functions of the Xs, such as restricted cubic spline functions for continuous Xs or a series of dummy variables for polytomous Xs. The expansion of each X_j is regressed separately on PC_1.
3. These separately fitted regressions specify the working transformations of each X.
4. Recompute PC_1 by doing a PC analysis on the transformed Xs (predicted values from the fits).
5. Repeat steps 2 to 4 until the proportion of variation explained by PC_1 reaches a plateau. This typically requires three to four iterations.

A transformation procedure that is similar to MTV is the maximum generalized variance (MGV) method due to Sarle [368, pp. 1267–1268]. MGV involves predicting each variable from (the current transformations of) all the other variables. When predicting variable i, that variable is represented as a set of linear and nonlinear terms (e.g., spline components). Analysis of canonical variates[279] can be used to find the linear combination of terms for X_i (i.e., find a new transformation for X_i) and the linear combination of the current transformations of all other variables (representing each variable as a single, transformed, variable) such that these two linear combinations have maximum correlation. (For example, if there are only two variables X_1 and X_2 represented as quadratic polynomials, solve for a, b, c, d such that $aX_1 + bX_1^2$ has maximum correlation with $cX_2 + dX_2^2$.) The process is repeated until the transformations converge. The goal of MGV is to transform each variable so that it is most similar to predictions from the other transformed variables. MGV does not use PCs (so one need not precede the analysis by variable clustering), but once all variables have been transformed, you may want to summarize them with the first PC.

The SAS `PRINQUAL` procedure of Kuhfeld[368] implements the MTV and MGV methods, and allows for very flexible transformations of the predictors, including monotonic splines and ordinary cubic splines.

A very flexible automatic procedure for transforming each predictor in turn, based on all remaining predictors, is the ACE (alternating conditional expectation) procedure of Breiman and Friedman.[68] Like SAS PROC `PRINQUAL`, ACE handles monotonically restricted transformations and categorical variables. It fits transformations by maximizing R^2 between one variable and a set of variables. It automatically transforms all variables, using the "super smoother"[207] for continuous variables. Unfortunately, ACE does not handle missing values. See Chapter 16 for more about ACE.

It must be noted that at best these automatic transformation procedures generally find only *marginal* transformations, not transformations of each predictor adjusted for the effects of all other predictors. When adjusted transformations differ markedly from marginal transformations, only joint modeling of all predictors (and the response) will find the correct transformations.

Once transformations are estimated using only predictor information, the adequacy of each predictor's transformation can be checked by graphical methods, by nonparametric smooths of transformed X_j versus Y, or by expanding the transformed X_j using a spline function. This approach of checking that transformations are optimal with respect to Y uses the response data, but it accepts the initial transformations unless they are significantly inadequate. If the sample size is low, or if PC_1 for the group of variables used in deriving the transformations is deemed an adequate summary of those variables, that PC_1 can be used in modeling. In that way, data reduction is accomplished two ways: by not using Y to estimate multiple coefficients for a single predictor, and by reducing related variables into a single score, after transforming them. See Chapter 8 for a detailed example of these scaling techniques.

4.7.4 Simultaneous Transformation and Imputation

As mentioned in Chapter 3 (p. 52) if transformations are complex or nonmonotonic, ordinary imputation models may not work. SAS PROC `PRINQUAL` implemented a method for simultaneously imputing missing values while solving for transformations. Unfortunately, the imputation procedure frequently converges to imputed values that are outside the allowable range of the data. This problem is more likely when multiple variables are missing on the same subjects, since the transformation algorithm may simply separate missings and nonmissings into clusters.

A simple modification of the MGV algorithm of `PRINQUAL` that simultaneously imputes missing values without these problems is implemented in the `R` function `transcan`. Imputed values are initialized to medians of continuous variables and the most frequent category of categorical variables. For continuous variables, transformations are initialized to linear functions. For categorical ones, transformations may be initialized to the identify function, to dummy variables indicating whether the observation has the most prevalent categorical value, or to random numbers. Then when using canonical variates to transform each variable in turn, observations that are missing on the current "dependent" variable are excluded from consideration, although missing values for the current set of "predictors" are imputed. Transformed variables are normalized to have mean 0 and standard deviation 1. Although categorical variables are scored using the first canonical variate, `transcan` has an option to use recursive partitioning to obtain imputed values on the original scale (Section 2.5) for these variables. It defaults to imputing categorical variables using the category whose predicted canonical score is closest to the predicted score.

`transcan` uses restricted cubic splines to model continuous variables. It does not implement monotonicity constraints. `transcan` automatically constrains imputed values (both on transformed and original scales) to be in the same range as non-imputed ones. This adds much stability to the resulting estimates although it can result in a boundary effect. Also, imputed values can optionally be shrunken using Eq. 4.5 to avoid overfitting when developing the imputation models. Optionally, missing values can be set to specified constants rather than estimating them. These constants are ignored during the transformation-estimation phase[n]. This technique has proved to be helpful when, for example, a laboratory test is not ordered because a physician thinks the patient has returned to normal with respect to the lab parameter measured by the test. In that case, it's better to use a normal lab value for missings.

The transformation and imputation information created by `transcan` may be used to transform/impute variables in datasets not used to develop the transformation and imputation formulas. There is also an R function to create R functions that compute the final transformed values of each predictor given input values on the original scale.

As an example of non-monotonic transformation and imputation, consider a sample of 1000 hospitalized patients from the SUPPORT[o] study.[352] Two mean arterial blood pressure measurements were set to missing.

```
require(Hmisc)
getHdata(support)   # Get data frame from web site
heart.rate      ← support$hrt
blood.pressure ← support$meanbp
blood.pressure[400:401]
```

```
Mean Arterial Blood Pressure Day 3
[1] 151 136
```

```
blood.pressure[400:401] ← NA   # Create two missings
d ← data.frame(heart.rate, blood.pressure)
par(pch=46)   # Figure 4.3
w ← transcan(∼ heart.rate + blood.pressure, transformed=TRUE,
             imputed=TRUE, show.na=TRUE, data=d)
```

```
Convergence criterion:2.901 0.035
```

```
0.007
Convergence in 4 iterations
$R^2$ achieved in predicting each variable:

    heart.rate blood.pressure
         0.259          0.259

Adjusted $R^2$:

    heart.rate blood.pressure
         0.254          0.253
```

[n] If one were to estimate transformations without removing observations that had these constants inserted for the current Y-variable, the resulting transformations would likely have a spike at $Y =$ imputation constant.

[o] Study to Understand Prognoses Preferences Outcomes and Risks of Treatments

```
w$imputed$blood.pressure
```

```
     400      401
132.4057 109.7741
```

```
t ← w$transformed
spe ← round(c(spearman(heart.rate, blood.pressure),
              spearman(t[,'heart.rate'],
                       t[,'blood.pressure'])), 2)
```

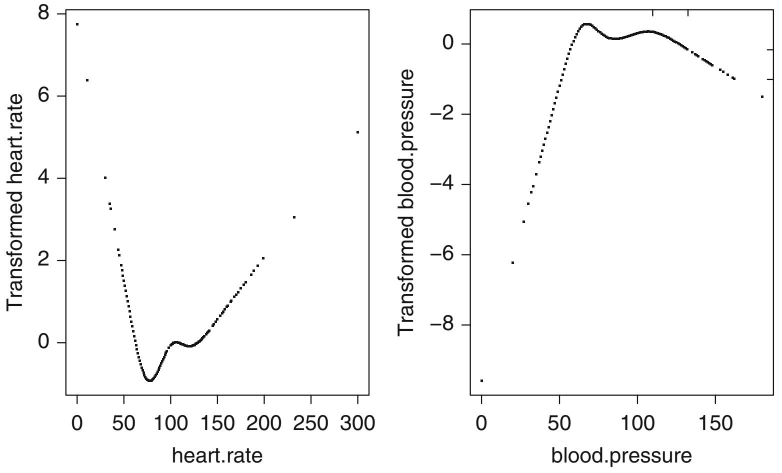

Fig. 4.3 Transformations fitted using `transcan`. Tick marks indicate the two imputed values for blood pressure.

```
plot(heart.rate, blood.pressure)   # Figure 4.4
plot(t[,'heart.rate'], t[,'blood.pressure'],
     xlab='Transformed hr', ylab='Transformed bp')
```

Spearman's rank correlation ρ between pairs of heart rate and blood pressure was -0.02, because these variables each require U-shaped transformations. Using restricted cubic splines with five knots placed at default quantiles, `transcan` provided the transformations shown in Figure 4.3. Correlation between transformed variables is $\rho = -0.13$. The fitted transformations are similar to those obtained from relating these two variables to time until death.

4.7.5 Simple Scoring of Variable Clusters

If a subset of the predictors is a series of related dichotomous variables, a simpler data reduction strategy is sometimes employed. First, construct two

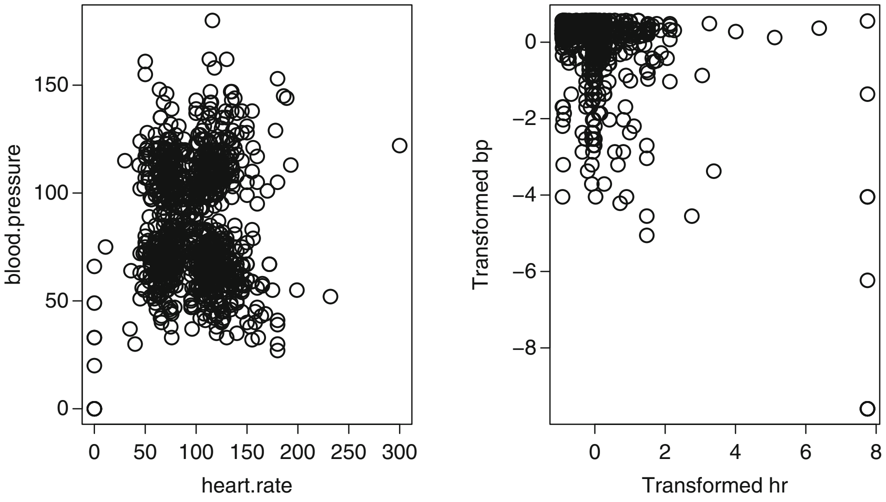

Fig. 4.4 The lower left plot contains raw data (Spearman $\rho = -0.02$); the lower right is a scatterplot of the corresponding transformed values ($\rho = -0.13$). Data courtesy of the SUPPORT study[352].

new predictors representing whether any of the factors is positive and a count of the number of positive factors. For the ordinal count of the number of positive factors, score the summary variable to satisfy linearity assumptions as discussed previously. For the more powerful predictor of the two summary measures, test for adequacy of scoring by using all dichotomous variables as candidate predictors after adjusting for the new summary variable. A residual χ^2 statistic can be used to test whether the summary variable adequately captures the predictive information of the series of binary predictors.[p] This statistic will have degrees of freedom equal to one less than the number of binary predictors when testing for adequacy of the summary count (and hence will have low power when there are many predictors). Stratification by the summary score and examination of responses over cells can be used to suggest a transformation on the score.

Another approach to scoring a series of related dichotomous predictors is to have "experts" assign severity points to each condition and then to either sum these points or use a hierarchical rule that scores according to the condition with the highest points (see Section 14.3 for an example). The latter has the advantage of being easy to implement for field use. The adequacy of either type of scoring can be checked using tests of linearity in a regression model[q].

[p] Whether this statistic should be used to change the model is problematic in view of model uncertainty.

[q] The R function `score.binary` in the `Hmisc` package (see Section 6.2) assists in computing a summary variable from the series of binary conditions.

4.7.6 Simplifying Cluster Scores

If a variable cluster contains many individual predictors, parsimony may sometimes be achieved by predicting the cluster score from a subset of its components (using linear regression or CART (Section 2.5), for example). Then a new cluster score is created and the response model is rerun with the new score in the place of the original one. If one constituent variable has a very high R^2 in predicting the original cluster score, the single variable may sometimes be substituted for the cluster score in refitting the model without loss of predictive discrimination. 22

Sometimes it may be desired to simplify a variable cluster by asking the question "which variables in the cluster are really the predictive ones?," even though this approach will usually cause true predictive discrimination to suffer. For clusters that are retained after limited step-down modeling, the entire list of variables can be used as candidate predictors and the step-down process repeated. All variables contained in clusters that were not selected initially are ignored. A fair way to validate such two-stage models is to use a resampling method (Section 5.3) with scores for deleted clusters as candidate variables for each resample, along with all the individual variables in the clusters the analyst really wants to retain. A method called *battery reduction* can be used to delete variables from clusters by determining if a subset of the variables can explain most of the variance explained by PC_1 (see [142, Chapter 12] and[445]). This approach does not require examination of associations with Y. Battery reduction can also be used to find a set of individual variables that capture much of the information in the first k principal components. 23

4.7.7 How Much Data Reduction Is Necessary?

In addition to using the sample size to degrees of freedom ratio as a rough guide to how much data reduction to do before model fitting, the heuristic shrinkage estimate in Equation 4.3 can also be informative. First, fit a full model with all candidate variables, nonlinear terms, and hypothesized interactions. Let p denote the number of parameters in this model, aside from any intercepts. Let LR denote the log likelihood ratio χ^2 for this full model. The estimated shrinkage is $(\mathrm{LR} - p)/\mathrm{LR}$. If this falls below 0.9, for example, we may be concerned with the lack of calibration the model may experience on new data. Either a shrunken estimator or data reduction is needed. A reduced model may have acceptable calibration if associations with Y are not used to reduce the predictors.

A simple method, with an assumption, can be used to estimate the target number of total regression degrees of freedom q in the model. In a "best case," the variables removed to arrive at the reduced model would have no association with Y. The expected value of the χ^2 statistic for testing those variables would then be $p - q$. The shrinkage for the reduced model is then on average $[\mathrm{LR} - (p - q) - q]/[\mathrm{LR} - (p - q)]$. Setting this ratio to be ≥ 0.9 and solving for q gives $q \leq (\mathrm{LR} - p)/9$. Therefore, reduction of dimensionality down to q degrees of freedom would be expected to achieve $< 10\%$ shrinkage. With these assumptions, there is no hope that a reduced model would have acceptable calibration unless $\mathrm{LR} > p + 9$. If the information explained by the omitted variables is less than one would expect by chance (e.g., their total χ^2 is extremely small), a reduced model could still be beneficial, as long as the conservative bound $(\mathrm{LR} - q)/\mathrm{LR} \geq 0.9$ or $q \leq \mathrm{LR}/10$ were achieved. This conservative bound assumes that no χ^2 is lost by the reduction, that is that the final model $\chi^2 \approx \mathrm{LR}$. This is unlikely in practice. Had the $p - q$ omitted variables had a larger χ^2 of $2(p - q)$ (the break-even point for AIC), q must be $\leq (LR - 2p)/8$.

As an example, suppose that a binary logistic model is being developed from a sample containing 45 events on 150 subjects. The 10:1 rule suggests we can analyze 4.5 degrees of freedom. The analyst wishes to analyze age, sex, and 10 other variables. It is not known whether interaction between age and sex exists, and whether age is linear. A restricted cubic spline is fitted with four knots, and a linear interaction is allowed between age and sex. These two variables then need $3 + 1 + 1 = 5$ degrees of freedom. The other 10 variables are assumed to be linear and to not interact with themselves or age and sex. There is a total of 15 d.f. The full model with 15 d.f. has $\mathrm{LR} = 50$. Expected shrinkage from this model is $(50 - 15)/50 = 0.7$. Since $\mathrm{LR} > 15 + 9 = 24$, some reduction might yield a better validating model. Reduction to $q = (50 - 15)/9 \approx 4$ d.f. would be necessary, assuming the reduced LR is about $50 - (15 - 4) = 39$. In this case the 10:1 rule yields about the same value for q. The analyst may be forced to assume that age is linear, modeling 3 d.f. for age and sex. The other 10 variables would have to be reduced to a single variable using principal components or another scaling technique. The AIC-based calculation yields a maximum of 2.5 d.f.

If the goal of the analysis is to make a series of hypothesis tests (adjusting P-values for multiple comparisons) instead of to predict future responses, the full model would have to be used.

A summary of the various data reduction methods is given in Figure 4.5.

When principal component analysis or related methods are used for data reduction, the model may be harder to describe since internal coefficients are "hidden." R code on p. 141 shows how an ordinary linear model fit can be used in conjunction with a logistic model fit based on principal components to draw a nomogram with axes for all predictors. 24

Fig. 4.5 Summary of Some Data Reduction Methods

Goals	Reasons	Methods
Group predictors so that each group represents a single dimension that can be summarized with a single score	• ↓ d.f. arising from multiple predictors • Make PC_1 more reasonable summary	Variable clustering • Subject matter knowledge • Group predictors to maximize proportion of variance explained by PC_1 of each group • Hierarchical clustering using a matrix of similarity measures between predictors
Transform predictors	• ↓ d.f. due to nonlinear and dummy variable components • Allows predictors to be optimally combined • Make PC_1 more reasonable summary • Use in customized model for imputing missing values on each predictor	• Maximum total variance on a group of related predictors • Canonical variates on the total set of predictors
Score a group of predictors	↓ d.f. for group to unity	• PC_1 • Simple point scores
Multiple dimensional scoring of all predictors	↓ d.f. for all predictors combined	Principal components $1, 2, \ldots, k, k < p$ computed from all transformed predictors

4.8 Other Approaches to Predictive Modeling

The approaches recommended in this text are

- fitting fully pre-specified models without deletion of “insignificant” predictors
- using data reduction methods (masked to Y) to reduce the dimensionality of the predictors and then fitting the number of parameters the data’s information content can support

- 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. 25

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 and median absolute errors are excellent measures for judging the value of a model developed without transforming the response to a model fitted after transforming Y, then back-transforming to get predictions. 26

4.11 Improving the Practice of Multivariable Prediction

Standards for published predictive modeling and feature selection in high-dimensional problems are not very high. There are several things that a good analyst can do to improve the situation.

1. Insist on validation of predictive models and discoveries, using rigorous internal validation based on resampling or using external validation.
2. Show collaborators that split-sample validation is not appropriate unless the number of subjects is huge
   - This can be demonstrated by spliting the data more than once and seeing volatile results, and by calculating a confidence interval for the predictive accuracy in the test dataset and showing that it is very wide.
3. Run a simulation study with no real associations and show that associations are easy to find if a dangerous data mining procedure is used. Alternately, analyze the collaborator's data after randomly permuting the Y vector and show some "positive" findings.
4. Show that alternative explanations are easy to posit. For example:
   - The importance of a risk factor may disappear if 5 "unimportant" risk factors are added back to the model
   - Omitted main effects can explain away apparent interactions.
   - Perform a *uniqueness analysis*: attempt to predict the predicted values from a model derived by data torture from all of the features not used in the model. If one can obtain $R^2 = 0.85$ in predicting the "winning" feature signature (predicted values) from the "losing" features, the "winning" pattern is not unique and may be unreliable.

4.12 Summary: Possible Modeling Strategies

Some possible global modeling strategies are to

- Use a method known not to work well (e.g., stepwise variable selection without penalization; recursive partitioning resulting in a single tree), document how poorly the model performs (e.g. using the bootstrap), and use the model anyway
- Develop a black box model that performs poorly and is difficult to interpret (e.g., does not incorporate penalization)

- Develop a black box model that performs well and is difficult to interpret
- Develop interpretable approximations to the black box
- Develop an interpretable model (e.g. give priority to additive effects) that performs well and is likely to perform equally well on future data from the same stream.

As stated in the Preface, the strategy emphasized in this text, stemming from the last philosophy, is to decide how many degrees of freedom can be "spent," where they should be spent, and then to spend them. If statistical tests or confidence limits are required, later reconsideration of how d.f. are spent is not usually recommended. In what follows some default strategies are elaborated. These strategies are far from failsafe, but they should allow the reader to develop a strategy that is tailored to a particular problem. At the least these default strategies are concrete enough to be criticized so that statisticians can devise better ones.

4.12.1 Developing Predictive Models

The following strategy is generic although it is aimed principally at the development of accurate predictive models.

1. Assemble as much accurate pertinent data as possible, with wide distributions for predictor values. For survival time data, follow-up must be sufficient to capture enough events as well as the clinically meaningful phases if dealing with a chronic process.
2. Formulate good hypotheses that lead to specification of relevant candidate predictors and possible interactions. Don't use Y (either informally using graphs, descriptive statistics, or tables, or formally using hypothesis tests or estimates of effects such as odds ratios) in devising the list of candidate predictors.
3. If there are missing Y values on a small fraction of the subjects but Y can be reliably substituted by a surrogate response, use the surrogate to replace the missing values. Characterize tendencies for Y to be missing using, for example, recursive partitioning or binary logistic regression. Depending on the model used, even the information on X for observations with missing Y can be used to improve precision of $\hat{\beta}$, so multiple imputation of Y can sometimes be effective. Otherwise, discard observations having missing Y.
4. Impute missing Xs if the fraction of observations with any missing Xs is not tiny. Characterize observations that had to be discarded. Special imputation models may be needed if a continuous X needs a non-monotonic transformation (p. 52). These models can simultaneously impute missing values while determining transformations. In most cases, multiply impute missing Xs based on other Xs and Y, and other available information about the missing data mechanism.

5. For each predictor specify the complexity or degree of nonlinearity that should be allowed (see Section 4.1). When prior knowledge does not indicate that a predictor has a linear effect on the property $C(Y|X)$ (the property of the response that *can* be linearly related to X), specify the number of degrees of freedom that should be devoted to the predictor. The d.f. (or number of knots) can be larger when the predictor is thought to be more important in predicting Y or when the sample size is large.
6. If the number of terms fitted or tested in the modeling process (counting nonlinear and cross-product terms) is too large in comparison with the number of outcomes in the sample, use data reduction (ignoring Y) until the number of remaining free variables needing regression coefficients is tolerable. Use the $m/10$ or $m/15$ rule or an estimate of likely shrinkage or overfitting (Section 4.7) as a guide. Transformations determined from the previous step may be used to reduce each predictor into 1 d.f., or the transformed variables may be clustered into highly correlated groups if more data reduction is required. Alternatively, use penalized estimation with the entire set of variables. This will also effectively reduce the total degrees of freedom.[272]
7. Use the entire sample in the model development as data are too precious to waste. If steps listed below are too difficult to repeat for each bootstrap or cross-validation sample, hold out test data from **all** model development steps that follow.
8. When you can test for model complexity in a very structured way, you may be able to simplify the model without a great need to penalize the final model for having made this initial look. For example, it can be advisable to test an entire group of variables (e.g., those more expensive to collect) and to either delete or retain the entire group for further modeling, based on a single P-value (especially if the P value is not between 0.05 and 0.2). Another example of structured testing to simplify the “initial” model is making *all* continuous predictors have the same number of knots k, varying k from 0 (linear), $3, 4, 5, \ldots$, and choosing the value of k that optimizes AIC. A composite test of all nonlinear effects in a model can also be used, and statistical inferences are not invalidated if the global test of nonlinearity yields $P > 0.2$ or so and the analyst deletes all nonlinear terms.
9. Make tests of linearity of effects in the model only to demonstrate to others that such effects are often statistically significant. Don’t remove insignificant effects from the model when tested separately by predictor. Any examination of the response that might result in simplifying the model needs to be accounted for in computing confidence limits and other statistics. It is preferable to retain the complexity that was prespecified in Step 5 regardless of the results of assessments of nonlinearity.

10. Check additivity assumptions by testing prespecified interaction terms. If the global test for additivity is significant or equivocal, all prespecified interactions should be retained in the model. If the test is decisive (e.g., $P > 0.3$), all interaction terms can be omitted, and in all likelihood there is no need to repeat this pooled test for each resample during model validation. In other words, one can assume that had the global interaction test been carried out for each bootstrap resample it would have been insignificant at the 0.05 level more than, say, 0.9 of the time. In this large P-value case the pooled interaction test did not induce an uncertainty in model selection that needed accounting.
11. Check to see if there are overly influential observations.
12. Check distributional assumptions and choose a different model if needed.
13. Do limited backwards step-down variable selection if parsimony is more important than accuracy.[582] The cost of doing any aggressive variable selection is that the variable selection algorithm must also be included in a resampling procedure to properly validate the model or to compute confidence limits and the like.
14. This is the "final" model.
15. Interpret the model graphically (Section 5.1) and by examining predicted values and using appropriate significance tests without trying to interpret some of the individual model parameters. For collinear predictors obtain pooled tests of association so that competition among variables will not give misleading impressions of their total significance.
16. Validate the final model for calibration and discrimination ability, preferably using bootstrapping (see Section 5.3). Steps 9 to 13 must be repeated for each bootstrap sample, at least approximately. For example, if age was transformed when building the final model, and the transformation was suggested by the data using a fit involving age and age^2, each bootstrap repetition should include both age variables with a possible step-down from the quadratic to the linear model based on automatic significance testing at each step.
17. Shrink parameter estimates if there is overfitting but no further data reduction is desired, if shrinkage was not built into the estimation process.
18. When missing values were imputed, adjust final variance–covariance matrix for imputation wherever possible (e.g., using bootstrap or multiple imputation). This may affect some of the other results.
19. When all steps of the modeling strategy can be automated, consider using Faraway's method[186] to penalize for the randomness inherent in the multiple steps. 27
20. Develop simplifications to the full model by approximating it to any desired degrees of accuracy (Section 5.5).

4.12.2 Developing Models for Effect Estimation

By effect estimation is meant point and interval estimation of differences in properties of the responses between two or more settings of some predictors, or estimating some function of these differences such as the antilog. In ordinary multiple regression with no transformation of Y such differences are absolute estimates. In regression involving $\log(Y)$ or in logistic or proportional hazards models, effect estimation is, at least initially, concerned with estimation of relative effects. As discussed on pp. 4 and 224, estimation of absolute effects for these models must involve accurate prediction of overall response values, so the strategy in the previous section applies.

When estimating differences or relative effects, the bias in the effect estimate, besides being influenced by the study design, is related to how well subject heterogeneity and confounding are taken into account. The variance of the effect estimate is related to the distribution of the variable whose levels are being compared, and, in least squares estimates, to the amount of variation "explained" by the entire set of predictors. Variance of the estimated difference can increase if there is overfitting. So for estimation, the previous strategy largely applies.

The following are differences in the modeling strategy when effect estimation is the goal.

1. There is even less gain from having a parsimonious model than when developing overall predictive models, as estimation is usually done at the time of analysis. Leaving insignificant predictors in the model increases the likelihood that the confidence interval for the effect of interest has the stated coverage. By contrast, overall predictions are conditional on the values of all predictors in the model. The variance of such predictions is increased by the presence of unimportant variables, as predictions are still conditional on the particular values of these variables (Section 5.5.1) and cancellation of terms (which occurs when differences are of interest) does not occur.
2. Careful consideration of inclusion of interactions is still a major consideration for estimation. If a predictor whose effects are of major interest is allowed to interact with one or more other predictors, effect estimates must be conditional on the values of the other predictors and hence have higher variance.
3. A major goal of imputation is to avoid lowering the sample size because of missing values in adjustment variables. If the predictor of interest is the only variable having a substantial number of missing values, multiple imputation is less worthwhile, unless it corrects for a substantial bias caused by deletion of nonrandomly missing data.

4. The analyst need not be very concerned about conserving degrees of freedom devoted to the predictor of interest. The complexity allowed for this variable is usually determined by prior beliefs, with compromises that consider the bias-variance trade-off.
5. If penalized estimation is used, the analyst may wish to not shrink parameter estimates for the predictor of interest.
6. Model validation is not necessary unless the analyst wishes to use it to quantify the degree of overfitting.

4.12.3 Developing Models for Hypothesis Testing

A default strategy for developing a multivariable model that is to be used as a basis for hypothesis testing is almost the same as the strategy used for estimation.

1. There is little concern for parsimony. A full model fit, including insignificant variables, will result in more accurate P-values for tests for the variables of interest.
2. Careful consideration of inclusion of interactions is still a major consideration for hypothesis testing. If one or more predictors interacts with a variable of interest, either separate hypothesis tests are carried out over the levels of the interacting factors, or a combined “main effect + interaction” test is performed. For example, a very well–defined test is whether treatment is effective for *any* race group.
3. If the predictor of interest is the only variable having a substantial number of missing values, multiple imputation is less worthwhile. In some cases, multiple imputation may increase power (e.g., in ordinary multiple regression one can obtain larger degrees of freedom for error) but in others there will be little net gain. However, the test can be biased due to exclusion of nonrandomly missing observations if imputation is not done.
4. As before, the analyst need not be very concerned about conserving degrees of freedom devoted to the predictor of interest. The degrees of freedom allowed for this variable is usually determined by prior beliefs, with careful consideration of the trade-off between bias and power.
5. If penalized estimation is used, the analyst should not shrink parameter estimates for the predictors being tested.
6. Model validation is not necessary unless the analyst wishes to use it to quantify the degree of overfitting. This may shed light on whether there is overadjustment for confounders.

4.13 Further Reading

1 Some good general references that address modeling strategies are [216,269,476,590].

2 Even though they used a generalized correlation index for *screening* variables and not for transforming them, Hall and Miller[249] present a related idea, computing the ordinary R^2 against a cubic spline transformation of each potential predictor.

3 Simulation studies are needed to determine the effects of modifying the model based on assessments of "predictor promise." Although it is unlikely that this strategy will result in regression coefficients that are biased high in absolute value, it may on some occasions result in somewhat optimistic standard errors and a slight elevation in type I error probability. Some simulation results may be found on the Web site. Initial promising findings for least squares models for two uncorrelated predictors indicate that the procedure is conservative in its estimation of σ^2 and in preserving type I error.

4 Verweij and van Houwelingen[640] and Shao[565] describe how cross-validation can be used in formulating a stopping rule. Luo et al.[430] developed an approach to tuning forward selection by adding noise to Y.

5 Roecker[528] compared forward variable selection (FS) and all possible subsets selection (APS) with full model fits in ordinary least squares. APS had a greater tendency to select smaller, less accurate models than FS. Neither selection technique was as accurate as the full model fit unless more than half of the candidate variables was redundant or unnecessary.

6 Wiegand[668] showed that it is not very fruitful to try different stepwise algorithms and then to be comforted by agreements in some of the variables selected. It is easy for different stepwise methods to agree on the wrong set of variables.

7 Other results on how variable selection affects inference may be found in Hurvich and Tsai[316] and Breiman [66, Section 8.1].

8 Goring et al.[227] presented an interesting analysis of the huge bias caused by conditioning analyses on statistical significance in a high-dimensional genetics context.

9 Steyerberg et al.[589] have comparisons of smoothly penalized estimators with the lasso and with several stepwise variable selection algorithms.

10 See Weiss,[656] Faraway,[186] and Chatfield[100] for more discussions of the effect of not prespecifying models, for example, dependence of point estimates of effects on the variables used for adjustment.

11 Greenland[241] provides an example in which overfitting a logistic model resulted in far too many predictors with $P < 0.05$.

12 See Peduzzi et al.[486,487] for studies of the relationship between "events per variable" and types I and II error, accuracy of variance estimates, and accuracy of normal approximations for regression coefficient estimators. Their findings are consistent with those given in the text (but[644] has a slightly different take). van der Ploeg et al.[629] did extensive simulations to determine the events per variable ratio needed to avoid a drop-off (in an independent test sample) in more than 0.01 in the *c*-index, for a variety of predictive methods. They concluded that support vector machines, neural networks, and random forests needed far more events per variable to achieve freedom from overfitting than does logistic regression, and that recursive partitioning was not competitive. Logistic regression required between 20 and 50 events per variable to avoid overfitting. Different results might have been obtained had the authors used a proper accuracy score.

13 Copas [122, Eq. 8.5] adds 2 to the numerator of Equation 4.3 (see also [504,631]).

14 An excellent discussion about such indexes may be found in http://r.789695.n4.nabble.com/Adjusted-R-squared-formula-in-lm-td4656857.html where J. Lucke points out that R^2 tends to $\frac{p}{n-1}$ when the population R^2 is zero, but R^2_{adj} converges to zero.

15 Efron [173, Eq. 4.23] and van Houwelingen and le Cessie[633] showed that the average expected optimism in a mean logarithmic quality score for a p-predictor binary logistic model is p/n. Taylor et al.[600] showed that the ratio of variances for certain quantities is proportional to the ratio of the number of parameters in two models. Copas stated that "Shrinkage can be particularly marked when stepwise fitting is used: the shrinkage is then closer to that expected of the full regression rather than of the subset regression actually fitted."[122, 504, 631] Spiegelhalter,[582] in arguing against variable selection, states that better prediction will often be obtained by fitting all candidate variables in the final model, shrinking the vector of regression coefficient estimates towards zero.

16 See Belsley [46, pp. 28–30] for some reservations about using VIF.

17 Friedman and Wall[208] discuss and provide graphical devices for explaining *suppression* by a predictor not correlated with the response but that is correlated with another predictor. Adjusting for a suppressor variable will increase the predictive discrimination of the model. Meinshausen[453] developed a novel hierarchical approach to gauging the importance of collinear predictors.

18 For incomplete principal component regression see [101, 119, 120, 142, 144, 320, 325]. See[396, 686] for sparse principal component analysis methods in which constraints are applied to loadings so that some of them are set to zero. The latter reference provides a principal component method for binary data. See[246] for a type of sparse principal component analysis that also encourages loadings to be similar for a group of highly correlated variables and allows for a type of variable clustering.See [390] for principal surfaces. Sliced inverse regression is described in [104, 119, 120, 189, 403, 404]. For material on variable clustering see [142, 144, 268, 441, 539]. A good general reference on cluster analysis is [634, Chapter 11]. de Leeuw and Mair in their R `homals` package [153] have one of the most general approaches to data reduction related to optimal scaling. Their approach includes nonlinear principal component analysis among several other multivariate analyses.

19 The redundancy analysis described here is related to *principal variables*[448] but is faster.

20 Meinshausen[453] developed a method of testing the importance of competing (collinear) variables using an interesting automatic clustering procedure.

21 The R `ClustOfVar` package by Marie Chavent, Vanessa Kuentz, Benoit Liquet, and Jerome Saracco generalizes variable clustering and explicitly handles a mixture of quantitative and categorical predictors. It also implements bootstrap cluster stability analysis.

22 Principal components are commonly used to summarize a cluster of variables. Vines[643] developed a method to constrain the principal component coefficients to be integers without much loss of explained variability.

23 Jolliffe[324] presented a way to discard some of the variables making up principal components. Wang and Gehan[649] presented a new method for finding subsets of predictors that approximate a set of principal components, and surveyed other methods for simplifying principal components.

24 See D'Agostino et al.[144] for excellent examples of variable clustering (including a two-stage approach) and other data reduction techniques using both statistical methods and subject-matter expertise.

25 Cook[118] and Pencina et al.[490, 492, 493] present an approach for judging the added value of new variables that is based on evaluating the extent to which the new information moves predicted probabilities higher for subjects having events and lower for subjects not having events. But see[292, 592].

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).