Chapter 10
Predictor Selection

Walter et al. (2001) developed a model to identify older adults at high risk of death in the first year after hospitalization, using data collected for 2,922 patients discharged from two hospitals in Ohio. Potential predictors included demographics, activities of daily living (ADLs), the APACHE-II illness-severity score, and information about the index hospitalization. A "backward" selection procedure with a restrictive inclusion criterion was used to choose a multipredictor model, using data from one of the two hospitals. The model was then validated using data from the other hospital. The goal was to select a model that best predicted future events, with a view toward identifying patients in need of more intensive monitoring and intervention.

Grodstein et al. (2001) evaluated the efficacy of hormone therapy (HT) for secondary prevention of CHD, using observational data for 2,489 women with a history of heart attack or documented coronary artery disease in the Nurse's Health Study (NHS), a prospective cohort followed from 1976 forward. In addition to measures of the use of HT, a set of known CHD risk factors were controlled for, including age, BMI, smoking, hypertension, LDL cholesterol levels, parental heart disease history, diet, and physical activity. The goal of predictor selection was to obtain a minimally confounded estimate of the effect of HT on risk of CHD events.

The Heart and Estrogen/Progestin Replacement Study (HERS), a randomized clinical trial addressing the same research question, was conducted among 2,763 postmenopausal women with clinically evident heart disease (Hulley et al. 1998). As in the NHS, a wide range of predictors were measured at study entry. Yet in the pre-specified analysis of the main HERS outcome, the only predictor was treatment assignment. The goal was to obtain a valid test of the null hypothesis as well as an unbiased estimate of the effectiveness of assignment to HT.

Orwoll et al. (1996) examined independent predictors of axial bone mass using data from the Study of Osteoporotic Fractures (SOF). SOF was a large ($n = 9{,}704$) observational cohort study designed to address multiple research questions about osteoporosis and fractures among ambulatory women aged 65 and up. Predictors considered by Orwoll had been identified in previous studies, and included weight, use of medications such as HT and diuretics, smoking history, alcohol and caffeine

E. Vittinghoff et al., *Regression Methods in Biostatistics*, Statistics for Biology and Health, DOI 10.1007/978-1-4614-1353-0_10,

use, calcium intake, physical activity, and various measures of physical function and strength. All variables that were statistically significant at $P < 0.05$ in models adjusting for age were included in the final multipredictor linear regression model. The goal was to identify all important predictors of bone mass.

In each of these examples, many more potential predictor variables had been measured than could reasonably be included in a multivariable regression model. The difficult problem of how to select predictors was resolved differently, to serve three distinct inferential goals:

(1) *Prediction*. Here, the primary issue is minimizing prediction error rather than causal interpretation of the predictors in the model. The prediction error of the model selected by Walter et al. (2001) was evaluated using an independent data set from a second hospital.
(2) *Evaluating a predictor of primary interest*. In pursuing this inferential goal, a central problem in observational data is confounding, which relatively inclusive models are more likely to minimize. Predictors necessary for *face validity* as well as those that behave like confounders should be included in the model. Randomized experiments like HERS represent a special case where the predictor of primary interest is the intervention; confounding is not usually an issue, but covariates are sometimes included in the model for other reasons.
(3) *Identifying the important independent predictors of an outcome*. This is the most difficult of the three inferential goals, and one in which both causal interpretation and statistical inference are most problematic. Pitfalls include false-positive associations, the potential complexity of causal pathways, and the difficulty of identifying a single best model. We also endorse inclusive models in this context, and recommend a selection procedure that affords increased protection against false-positive results. Cautious interpretation of weak associations is key to this approach.

In summary, *predictor selection* is the process of choosing appropriate predictors for inclusion in a multipredictor regression model. A good model should be substantively motivated, appropriate to the inferential goal and sample size, interpretable, and persuasive.

10.1 Prediction

In selecting a good prediction model, candidate predictors should be considered in terms of their contribution to reducing prediction error.

Definition: *Prediction error* (PE) measures how well the model is able to predict the outcome for new observations not used in developing the prediction model.

10.1.1 Bias–Variance Trade-off and Overfitting

Inclusive models that minimize confounding may not work as well for prediction as models with smaller numbers of predictors. This can be understood in terms of the *bias–variance trade-off*. Bias in predictions is often reduced when more variables are included in the model, provided they are measured and modeled adequately. Moreover, the coefficients are often nearly unbiased under the assumptions commonly made in these analyses. But as less important covariates are added to the model, precision may start to erode, without commensurate decreases in bias. The larger models may be *overfitted* to the idiosyncrasies of the data, and, thus, more poorly predict new, independent observations. We can minimize PE by optimizing the bias–variance trade-off.

10.1.2 Measures of Prediction Error

For continuous outcomes, R^2 is a potential measure of PE. A function of the residual sum of squares (RSS), R^2 depends on the averaged squared distance between the predictions, or fitted values, and the observed outcomes, and so is a natural metric for PE.

For binary outcomes, the analogous Brier score, also given by the average of the squared distances between the predicted and observed outcomes, is not commonly used. A much more widely used PE measure is the area under the ROC curve, or equivalently the C-statistic, introduced in Sect. 5.2.6. The analogous PE measure for Cox models is the C-index. The C-statistic and C-index are both measures of *discrimination*—that is, how effectively the model can distinguish between events and nonevents, or correctly order the timing of two events.

Both the C-statistic and C-index are rank-based measures, and can be insensitive to improvements in prediction as a result (Pencina et al. 2008). To see this, note that in calculating the C-statistic, two correctly ranked event/nonevent pairs for which the predictions differ by five and 95 percentage points would be treated alike, although the model much more clearly distinguishes the second pair. Likewise, in calculating the C-index, we ignore differences between failure times as well as between fitted risks.

In addition to discrimination, measures of *calibration* for logistic and Cox models assess the agreement between fitted and observed risks. The Hosmer–Lemeshow statistic presented in Chap. 5 measures calibration of the logistic model, comparing fitted and observed events within deciles (or other groupings) of the fitted risks. Analogs have been proposed for the Cox model (Parzen and Lipsitz 1999; van Houwelingen 2000). One often-used measure of calibration for the Cox model is to compare average fitted probabilities of an event within a fixed time period to observed probabilities nonparametrically estimated using Kaplan–Meier curves. For example, Cook et al. (2006) compared fitted and observed ten-year risks for cardiovascular events within two-point intervals of the model-based risk score.

10.1.3 Optimism-Corrected Estimates of Prediction Error

To select a model that minimizes prediction error, we need an accurate estimate of the target PE measure that does not overstate the ability of the model to predict the outcome for new, independent observations—in brief, one that is not *optimistic*.

10.1.3.1 Optimism of Naïve Estimates of PE

To see why optimism is an issue, consider R^2, the proportion of variance explained by a linear regression model, and a potential measure of PE. It increases with each additional covariate, even if the added predictor provides minimal information about the outcome. At the extreme, $R^2 = 1$ in a model with one predictor for each observation. This happens because the same observations are used to estimate the model and assess its predictiveness. Selecting predictors simply to maximize R^2 would almost surely result in overfitted models.

10.1.3.2 Simple Alternatives to R^2

An alternative less subject to optimism is adjusted R^2, which is calculated by penalizing R^2 for the number of predictors in the model. Thus, when a variable is added, adjusted R^2 increases only if the increment in R^2 is larger than the increment in the penalty. The Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) are analogs which impose stiffer penalties for each additional variable—specifically, penalties against minus twice the log-likelihood, another potential measure of PE. With AIC, the penalty is $2p$, where p is the number of predictors in the model; with BIC, it is $p \log N$, where N is the sample size.

The AIC criterion is relatively liberal, allowing for the inclusion of simple continuous or binary predictors with P-values <0.16. In contrast, the P-value cutoff imposed by BIC for such predictors grows progressively stricter with sample size, requiring $P < 0.05$ in samples of about 50, $P < 0.01$ in samples of 500, and $P < 0.009$ in samples of 1,000, and, thus, leads to increasingly parsimonious models, relative to AIC. Both measures depend on the number of additional coefficients, and so set the bar higher for inclusion of restricted cubic splines or multicategory predictors.

In Stata the `regress` command prints adjusted R^2 by default, and AIC and BIC can be obtained for linear, logistic, Cox, and other models using the postestimation command `estat ic`. The best prediction model is taken to be the one that maximizes adjusted R^2, or minimizes AIC or BIC.

10.1.3.3 Generalized Cross-Validation

In contrast to indirect, theoretically-based measures such as adjusted R^2, AIC, and BIC, more direct methods for obtaining nonoptimistic estimates of PE are based on *cross-validation*, which uses distinct, independent sets of observations to estimate the model and to evaluate PE.

10.1.3.4 Development and Validations Sets

The most straightforward example of cross-validation is the split-sample approach, in which the parameter estimates are obtained from a so-called development set, but then PE is evaluated in an independent validation set by comparing observed outcomes to expected values calculated using development set parameter estimates in combination with validation set covariate values.

In some implementations, the development and validation sets are obtained by splitting a single data set, often with two-thirds of the observations randomly assigned to the development set. Other implementations, as in Walter's analysis of posthospitalization mortality among high-risk older adults, use an independent sample as the validation set. Precisely because the validation set is not sampled under exactly the same circumstances, this procedure may do a better job of forecasting the utility of the prediction model in practical use. Altman and Royston (2000) discuss the merits of internal and external validation sets.

Splitting one data set into development and validation sets is less efficient than the alternative discussed next, but also easier to implement, and commonly more credible to nonstatisticians, in particular when the validation set is truly external.

10.1.3.5 *h*-Fold Cross-Validation

A more efficient alternative to splitting the data into development and validation sets is *h-fold cross-validation*. With this method, the entire data set is used both for development and validation of the model. The procedure works in five basic steps.

(1) The data are randomly divided into h mutually exclusive subsets of equal size.
(2) Each of the h subsets is set aside in turn, and the model is estimated using the remaining observations.
(3) Using the parameter estimates from each of those h models, the statistics necessary to calculate the target measure of PE are estimated for the corresponding set-aside observations.
(4) A summary estimate of PE is then calculated using the statistics from all h subsets.
(5) The h-fold procedure is repeated k times, using a new division of the data each time, and then the k summary estimates of PE are averaged.

Values of $h = 5-10$ and $k = 10-20$ are reasonable.

Table 10.1 Ten-fold cross-validation of the area under the ROC curve

```
. quietly logistic chd69 age chol sbp bmi smoke
. predict fitted, pr

. * Naive estimate of area under the ROC curve
. roctab chd69 fitted

                      ROC                    -Asymptotic Normal--
           Obs       Area     Std. Err.      [95% Conf. Interval]
         --------------------------------------------------------
          3142     0.7333       0.0156        0.70270     0.76395

. Step 1: divide data into 10 mutually exclusive subsets
. xtile group = uniform(), nq(10)
. gen cv_fitted = .
. forvalues i = 1/10 {
  2.
.         * Step 2: estimate model omitting each subset
             qui logistic ytemp age chol sbp bmi smoke if group~=`i'
  3.         qui predict cv_fittedi, pr
  4.
.         * Step 3: save cross-validated statistic for each omitted subset
.            qui replace cv_fitted = cv_fittedi if group==`i'
  5.         qui drop cv_fittedi
  6.         }

.
. * Step 4: calculate cross-validated area under ROC curve
. roctab chd69 cv_fitted

                      ROC                    -Asymptotic Normal--
           Obs       Area     Std. Err.      [95% Conf. Interval]
         --------------------------------------------------------
          3142     0.7277       0.0158        0.69386     0.75566
```

Cross-validation is easy to implement in Stata. In Table 10.1, we first re-run the logistic model for CHD risk shown in Table 5.6, save the fitted probabilities, and calculate the naïve estimate of the area under the ROC curve (`ROC Area`), equivalent to the *C*-statistic. Then, the WCGS data are randomly divided into ten mutually exclusive subsets, and the model is refitted ten times, omitting in turn each of the ten subsets from the data used in estimation of the model. However, predicted values are calculated for the entire data set; we also exploited this feature of Stata for potential outcomes estimation in Table 9.6. The cross-validation fitted values for the omitted subsets are collected in the new variable `cv_fitted`, and in a final step, the cross-validation estimate of the area under the ROC curve is calculated using these fitted values and the observed outcomes. For clarity, we have omitted the fifth step of repeating the procedure 10–20 times, but the additional programming is simple enough.

As expected, the optimistic naïve estimate of the area under the ROC curve shown in Table 10.1 is larger than the cross-validated estimate. However, the difference is small, suggesting that the simple logistic model for CHD events is not badly overfitted.

10.1.4 *Minimizing Prediction Error Without Overfitting*

A model that fits well, including all important predictors and accurately capturing nonlinear effects as well as interactions, should provide better prediction than a poorly specified model that excludes some important predictors, inaccurately models the effects of others, and includes unimportant predictors.

Earlier chapters have shown how to ensure that nonlinear effects of continuous predictors are adequately modeled, essentially by examining the relationship between predictor and outcome, using diagnostic plots or models including restricted cubic splines or interactions. And later in this chapter, in discussing predictor selection for the second inferential goal of evaluating the causal effect of a primary predictor of interest, we recommend methods to ensure that all measured confounders are included and adequately modeled, again by examining alternative models for the outcome.

However, in this context, the danger is that examining relationships with the outcome can easily lead to overfitting, resulting in a model that does not perform well in external validation data. Overfitting can be minimized using four strategies:

(1) Pre-specify well-motivated predictors and how to model them
(2) Eliminate predictors without using the outcome
(3) Use the outcome, but cross-validate the target measure of PE
(4) Use the outcome, and shrink the coefficient estimates.

10.1.4.1 Pre-specifying Well-Motivated Predictors

One primary strategy for avoiding overfitting is to depend so far as possible on a priori specification of well-motivated candidate predictors. In areas of clinical research where prognostic factors have been thoroughly studied, expert opinion, grounded in the literature, may provide considerable guidance, and meta-analyses can be especially reliable measures of variable importance. This strategy would also rely on the literature to determine how the effects of continuous covariates should be modeled—that is, to select functional form—rather than using the data to guide these decisions.

In some well-studied areas, this step may be sufficient to choose a good prediction model, without the need for subsequent elimination of predictors driven by the development data. Furthermore, while the bias–variance tradeoff may suggest the need for parsimony, a wisely-chosen set of pre-specified predictors may often work better in external validation data than a subset of those predictors chosen by looking at their relationships with the outcome in the data used for model development (Harrell 2005; Steyerberg 2009).

10.1.4.2 Predictor Elimination Without Using the Outcome

A second-line strategy for avoiding overfitting is to eliminate candidates without looking at predictor–outcome relationships, but taking account of the effective sample size m, defined as the number of observations in linear regression, the number of events in Cox regression, and the smaller of the numbers of observations with or without the outcome in logistic models (Harrell 2005).

For example, summary variables can be chosen for predictor domains: LDL and HDL cholesterol levels might be chosen on substantive grounds from among the larger set of lipid measures including total cholesterol, triglycerides, and the HDL/LDL ratio. Practical considerations may also be important. In particular, expensive, invasive, risky, and relatively unreliable tests can be ruled out if more practical alternatives are available. Predictors with fewer missing values in the development data are also preferable, in particular, if missing values reflect the likely difficulty of obtaining the measurement in practice.

Linearity would of course be a concern in modeling the effect of continuous covariates such as LDL cholesterol. To address this issue, a related means of outcome-free predictor elimination is to allocate spline knots based on prior estimates of variable importance and m. Thus, if a predictor has been of primary importance and had strongly nonlinear effects in earlier research, and m allows it, a four- or five-knot spline may be pre-specified. In contrast, a less important predictor or one known to have approximately linear effects can be treated more simply. Smaller samples and fewer outcomes may also limit how flexibly we can model continuous effects.

Principal components is a more complicated alternative for reducing the number of parameters to be estimated without using the outcome, and has been shown to work well in some studies (Harrell et al. 1984, 1996). This method summarizes a large set of correlated continuous predictors by a much smaller set of uncorrelated summary variables, or principal components, chosen to explain most of the variance in the predictors. This simplification is achieved without reference to the outcome.

This approach does have some drawbacks. One is that the principal components may not be substantively interpretable, which is desirable for face validity, although not really needed for prediction. In addition, principal components capturing the greatest variability in the *predictors* are not guaranteed to capture the most variability in the *outcome*, although with well-chosen predictors this is likely. Finally, this procedure does not reduce the number of underlying variables that need to be measured, and so makes it more difficult to focus on easily-obtained predictors with fewer missing values.

A widely used guideline suggests that at most $m/10$ or even $m/20$ candidate predictors should be *considered* for inclusion in the prediction model. Note that each component of a complicated predictor counts as an additional candidate, so that if we pre-specify a restricted cubic spline with five knots to represent a continuous predictor, the number of candidates increases by four, the required number of spline basis variables. Motivated by simulation studies of the precision of predictions based on Cox models, this guideline is approximate, but does suggest that large samples are necessary for developing valid prediction models, in particular, when variable selection is required.

10.1.4.3 Model Selection Using the Outcome and Cross-Validation

In the common case where the combination of prespecification and outcome-free predictor elimination does not adequately reduce the number of candidate predictors, an effective strategy is to use exhaustive screening of all subsets of the remaining candidate predictors. Crucially, to avoid overfitting, this final screening step must use cross-validation of a selected target measure of PE to help identify the most predictive of these models. This is the approach used in most modern algorithms for prediction model development, including the Deletion/Substitution/Addition (DSA) algorithm (Molinaro and van der Laan 2004). In this procedure, implemented in R, the candidate predictors, including polynomial terms and interactions, are efficiently screened using h-fold cross-validation of a selected measure of PE.

Efficient screening is an important issue in this context. For example, even if the number of candidate predictors has been reduced to a seemingly tractable eight, the number of subsets of all sizes is $2^8 = 256$. And even if an indirect optimism-corrected measure of prediction error—adjusted R^2, AIC, or BIC—is used in place of cross-validation, this represents an onerous computing task without programs like DSA that automate the screening.

However, screening can be made more practicable if some of the remaining candidates are always to be included on a priori grounds. For example, if five of eight candidate variables were to be included by default, then only $2^3 = 8$ models must be screened. But if many models have to be screened, programming of the procedure, including any intermediate steps, will almost surely be required. We illustrate this approach in Sect. 10.1.6 below.

While this screening procedure should help us find a good predictive model without overfitting, it is important to note that the cross-validated estimate of PE for the selected model will be at least slightly optimistic, not because we use the same data to estimate model parameters and evaluate PE—the source of optimism in naïve PE estimators—but because of the selection.

10.1.4.4 Shrinking the Coefficient Estimates

Dropping variables, on a priori or practical grounds or on the basis of a cross-validated PE measure, is equivalent to setting their coefficients equal to zero. An alternative approach is to shrink them only part way to zero. So-called *shrinkage* procedures can be motivated on the grounds that even the weaker candidate predictors specified a priori have some predictive value, and so should not be excluded outright from the model. However, because their coefficients may be less precisely estimated, better prediction may be achieved by reducing their influence. This approach is closely related to the shrinkage estimators introduced in Sect. 7.7.3.

In general shrinkage procedures impose penalties against the log-likelihood in model fitting, with the degree of penalization generally optimized using cross-validation. Le Cessie and Van Houwelingen (1992) and Verweij and Van Houwelingen (1994) discuss applications to logistic and Cox regression. These methods derive from *ridge regression* (Hoerl and Kennard 1970), which provides slightly biased but less variable estimates in linear models when the predictors are highly correlated. In ridge regression, the penalty is proportional to the sum of the squared values of the regression coefficients, with the proportionality factor commonly optimized using cross-validation. Coefficients are shrunken roughly in inverse proportion to the variance of the corresponding predictor, but no variables are omitted outright.

In contrast, the penalty imposed by the Least Absolute Shrinkage and Selection Operator (LASSO) (Tibshirani 1997) is proportional to the sum of the *absolute* values of the regression coefficients. Surprisingly, the result is that the LASSO can set the coefficients for the least important predictors to zero, effectively omitting those variables from the model, while differentially shrinking others. Thus, it is a selection as well as a shrinkage procedure. The LASSO has been implemented only for linear models in the Stata `lars` package, as so-called *least angle regression*. However, the `penalized` package in R extends both ridge regression and the LASSO to GLMs and Cox models, and incorporates cross-validation for selecting the penalty factor.

10.1.5 Point Scores

Unless a continuous predictor has strong threshold effects, we can generally achieve better prediction by keeping it continuous, modeling any nonlinearity in its effects, and avoiding dichotomization. However, one drawback, especially if splines are used to capture nonlinear effects, is that the predictions almost always need to be calculated using some electronic interface, or at least a nomogram. If the prediction model is intended for everyday clinical use, easily calculated scores assigning points to a small set of risk factors are more likely to be adopted.

For example, the Thrombosis in Myocardial Infarction (TIMI) risk score for predicting event-free survival in heart disease patients is simply calculated by counting up seven risk indications, including age ≥ 65, having ≥ 3 CAD risk factors, coronary stenosis, ST-segment deviation, elevated serum cardiac markers, ≥ 2 recent angina episodes, and aspirin use in the last week (Antman et al. 2000). Each of the underlying predictors was dichotomized and assigned one point.

At some cost in complexity, more information can be retained by splitting continuous variables into more than two categories, with nonreference levels assigned different numbers of points. For example, D'Agostino et al. (2000) tabulate points assigned to each level of several multicategory predictors, and provide an additional table for translating the summed point scores into predicted risks.

Point systems allowing differing weights are commonly derived by rounding the regression coefficients for each binary indicator variable, after suitable rescaling so that each factor is assigned at least one point. In some cases, risk scores of this type may perform nearly as well as summary scores based on the underlying coefficients. However, considerable increases in prediction error may sometimes result (Gordon et al. 2010).

10.1.6 Example: Risk Stratification of Patients with Heart Disease

The Heart and Soul Study follows a prospective cohort of 1,024 adults with established CHD, recruited from several clinical centers in the San Francisco Bay Area in 2000–2002 (Whooley et al. 2008). Over 5,745 person-years of follow-up by the time of analysis, 272 outcome events, a composite defined by heart attack, heart failure, stroke, or death from cardiovascular causes, had been observed among 916 of these participants with complete baseline test data.

Starting from a wide range of baseline predictors, we developed two Cox models for risk stratification of this moderate-to-high risk patient population. One, requiring computer implementation, includes three continuous predictors, two of them represented by 3-knot restricted cubic splines. The second is a point score model. We selected Harrell's C-index as our target PE measure, and drove final model selection mainly by minimizing cross-validated estimates of this target.

Based on the knowledge of the investigators, an initial set of 36 candidate predictors was identified. On practical grounds and by choosing—without using the outcomes—the best predictor in several domains, the number was reduced to 18, under the $m/10$ upper bound of 27, but still exceeding the more conservative bound of $m/20$. While cut-points for dichotomizing continuous predictors, as required for the point score, were available from the literature, less information was available on functional form. On practical grounds, the investigators specified that the point score should include at most 7 predictors, and preferably 5 or 6, and were reluctant to consider larger continuous models.

Because the number of possible models was very large even before considering the functional form of continuous predictors, we used exploratory analysis to reduce the scope of the cross-validation screening. Specifically, using backward selection procedures, we decided that the four clearly most powerful predictors (age, left ventricular ejection fraction (LVEF), B-natriuretic peptide levels (BNP), and urinary creatinine/albumin ratios (UACR), would be included in any selected model, and that we could safely omit the four weakest (hypertension, history of heart attack, LDL, and HDL cholesterol). The remaining candidates for inclusion in the model included gender, BMI, current smoking, diabetes, C-reactive protein (CRP), chronic kidney disease (CKD), detectable troponin, congestive heart failure (CHF), physical inactivity, and poor adherence to medication.

Table 10.2 Top-scoring prediction models

	Continuous			Point score		
Number of	*C*-Index (%)		GOF	*C*-Index (%)		GOF
Predictors	CV[a]	Naïve	*P*-value[b]	CV	Naïve	*P*-value
5	76.2	76.6	0.90	73.1	74.0	0.002
6	76.2	76.9	0.50	73.9	74.5	0.07
7	76.2	76.8	0.72	73.0	74.8	0.03

[a] Cross-validation.

[b] Goodness of fit test due to Parzen and Lipsitz (1999).

In addition, we selected the functional form for continuous predictors by comparing AIC values for alternatives, in models adjusting for other powerful covariates. On this basis, we elected to treat age as linear, dichotomized LVEF at the established cutpoint of 50%, and used 3-knot restricted cubic splines for UACR and BNP as well as BMI and CRP. In additional exploratory analyses using four or five-knot splines, the cross-validated *C*-index decreased substantially, reflecting overfitting.

We then programmed algorithms in Stata to perform 10-fold cross-validation of the *C*-index for each of several hundred candidate continuous and point score models. For the point-score models, we used a simple automatic algorithm for calculating the scores based on each of the ten cross-validation development sets.

Table 10.2 shows results for 5, 6, and 7-predictor continuous and point score models with values of the cross-validated *C*-index at the observed maximum. Several comments are in order:

- The continuous models consistently do better than the point score models. The 2%–3% point improvements in the *C*-index are substantial and not easily achieved. Note that the number of *parameters* estimated for the continuous models is two greater than the number of predictors, because BNP and UACR were modeled using 3-knot splines.
- The larger models have at most slightly higher cross-validated values of the *C*-index. Moreover, continuous models with more than 7 predictors did not do substantially better than the 7-predictor model.
- The naïve and cross-validated *C*-index values are also very close, possibly reflecting optimism of the cross-validated estimate due to selection.
- Holding the number of predictors or parameters fixed, the *C*-indices for the top 5–10 models barely differed (data not shown). This illustrates that in prediction, models containing different sets of predictors may be quite competitive.

When different models are close in terms of the cross-validated target measure of PE, additional criteria may be used to decide between them, including calibration. Despite the evidence for poor fit of the point score models, model-based and Kaplan–Meier estimates of risk were in reasonably good agreement for the 6-predictor models, as shown in Fig. 10.1, as well as for the 7-predictor and larger models. Results are stratified by decile of predicted risk for the continuous model, and by point scores for the point score model.

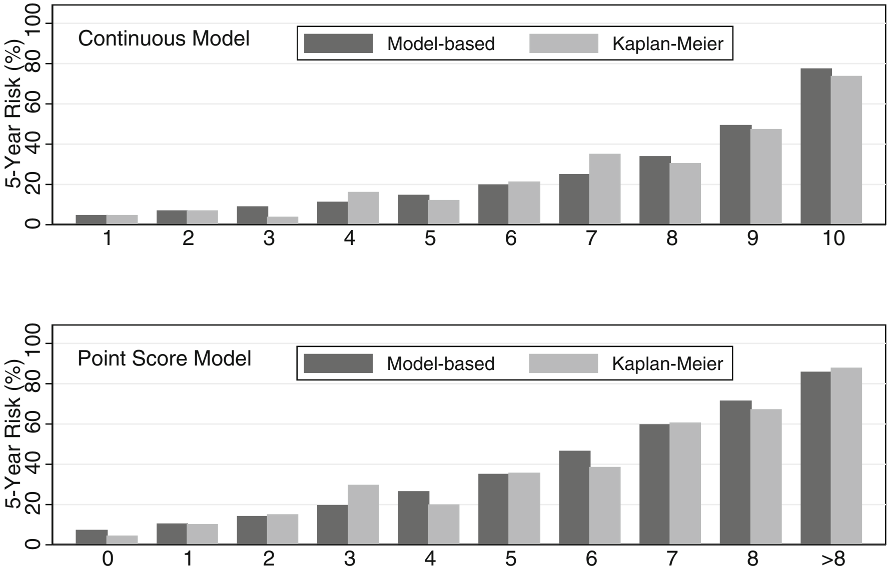

Fig. 10.1 Calibration of prediction models

Face validity and clinical convenience were also top priorities for the investigators. Current smoking and diabetes are accepted and reasonably strong cardiovascular risk factors. The best 7-predictor continuous models included either troponin or CRP, and so would have required an extra test, without improving discrimination or calibration. Accordingly, the investigators selected the 6-predictor model, including age, LVEF, BNP, UACR, current smoking, and diabetes.

10.2 Evaluating a Predictor of Primary Interest

In observational data, the main problem in evaluating a predictor of primary interest is to rule out confounding of the association between this predictor and the outcome as persuasively as possible. Potential confounders to be considered include factors identified in previous studies or hypothesized to matter on substantive grounds, as well as variables that behave like confounders by the statistical measures described in Sect. 4.4. Three classes of covariates would not be considered for inclusion in the model: covariates which are essentially alternative measures of either the outcome or the predictor of interest, and those hypothesized to mediate its effect. A diagram of the proposed causal model can be useful for clarifying hypotheses about these relationships, which can be complex, and for selecting variables for consideration.

In contrast, mediation of one confounder by another would not affect the estimate for the primary predictor nor its interpretation. Similarly, high correlation between pairs of adjustment of confounding variables would not necessarily be a compelling reason for removing one of them, if both are seen as necessary on substantive or statistical grounds; the reason is that collinearity between confounding variables will not affect the estimate for the primary predictor or its precision. Covariates which are in some sense alternative measures of the outcome are not always easy to recognize, but should usually be excluded. For example, it would be problematic to include diabetes in a model for glucose, because diabetes is largely defined by elevated glucose. Another example is history of a potentially recurrent outcome like falling in a model for subsequent incidence of the outcome. In both examples, addition of the alternative outcome measure as a predictor to the model tends to attenuate the estimates for other, more interpretable predictors.

10.2.1 Including Predictors for Face Validity

Some variables in the hypothesized causal model may be such well-established causal antecedents of the outcome that it makes sense to include them, essentially to establish the face validity of the model and without regard to the strength or statistical significance of their associations with the primary predictor and outcome in the current data set. The risk factors controlled for in the Nurse's Health Study analysis of the effects of HT on CHD risk are well understood and meet this criterion.

10.2.2 Selecting Predictors on Statistical Grounds

In many areas of research, the potential confounders of a predictor of interest may be less well established, so that in the common case where there are many such potential confounders, a priori selection of a reasonable subset to adjust for is not a realistic option. However, the inclusion of too many predictors may unacceptably inflate the standard errors of the regression coefficients, especially in smaller samples; in logistic and Cox models bias can also be induced when too many parameters are estimated. We discuss collinearity and the numbers of predictors that can safely be included in Sects. 10.4.1 and 10.4.2. Because of these potential problems, we would like to eliminate variables that are effectively not confounders, because they demonstrate little or no independent association with the outcome after adjustment. Similarly, hypothesized interactions that turn out not to be important on statistical grounds would be eliminated, almost always before either of the interacting main effects are removed.

An easily implemented method for eliminating redundant predictors on statistical grounds is so-called backward selection. In brief, backward selection begins with full model including all pre-specified candidate predictors, then sequentially eliminates the weaker candidates, at each step removing the predictor with the largest P-value. The advantages of backward over forward and stepwise procedures are explained in Sect. 10.4.3.

If P-value driven selection is used, we recommend a liberal criterion, to rule out confounding more effectively: in particular, only removing variables with P-values ≥ 0.2 (Maldonado and Greenland 1993). A comparably effective alternative is to retain variables if removing them changes the coefficient for the predictor of interest by more than 10% or 15% (Greenland 1989; Mickey and Greenland 1989). These liberal criteria are particularly important in small data sets, where even important confounders may not meet the usual $P < 0.05$ criterion for statistical significance.

10.2.3 Interactions With the Predictor of Primary Interest

A potentially important check on the validity of the selected model is to assess interactions between the primary predictor and important covariates, in particular, those that are biologically plausible. Especially for a novel or controversial main finding, it can add credibility to show that the association is similar across subgroups. There is no reason for concern if the association is statistically significant in one subgroup but not in the complementary group, provided the subgroup-specific estimates are similar. However, if a substantial and credible interaction is found, particularly such that the association with the predictor of interest differs qualitatively across subgroups, then the analysis would need to take account of this complexity. For example, Kanaya et al. (2004) found an interaction between change in obesity and HT in predicting CHD and mortality risk which substantively changed the interpretation of the finding. However, since such exploratory analyses are susceptible to false-positive findings, this unexpected and hard-to-explain interaction was cautiously interpreted.

10.2.4 Example: Incontinence as a Risk Factor for Falling

Brown et al. (2000) examined urinary incontinence as a risk factor for falling among 6,049 ambulatory, community-dwelling women in the SOF cohort also studied by Orwoll. The hypothesis was that incontinence might cause falling because of hasty trips to the bathroom, especially at night. But it was important to rule out confounding by physical decline, which is strongly associated with both aging and incontinence. The final model included all predictors which were associated with the outcome at $P < 0.2$ in univariable analysis and remained statistically significant at that level after multivariable adjustment. Alternative and more inclusive models with different sets of predictors were also assessed. After adjustment for 12 covariates (age; history of nonspine fracture and falling; living alone; physical activity; use of a cane, walker, or crutch; history of stroke or diabetes; use of two classes of drugs; a physical performance variable; and BMD) weekly or more frequent urge incontinence was independently associated with a 34% increase in risk of falling (95% CI 6%–69%, $P = 0.01$).

In this example, falling was defined as a binary outcome, discussed in Chap. 5. In addition, because the outcome was observed over multiple time intervals for each SOF participant, methods presented in Chap. 7 for longitudinal repeated measures were used. A subsequent example in Sect. 10.4.2 uses a Cox proportional hazards model, covered in Chap. 6. In using these varied examples, we underscore the fact that predictor selection issues are essentially the same for all the regression models covered in this book.

10.2.5 Directed Acyclic Graphs

So-called *directed acyclic graphs* (DAGs) (Pearl 1995), a type of causal diagram, are potentially useful in determining which covariates need to be included in—and excluded from—regression models used for the second inferential goal of evaluating the effects of a predictor of primary interest. In the following example we briefly review the terminology and some key ideas, show how a DAG could be used to guide predictor selection for this inferential goal, and discuss some complications that can arise.

10.2.5.1 Example: Vitamin Use and Birth Defects

Suppose we would like to assess the causal effect of vitamin use on prevention of birth defects. The DAG in Fig. 10.2 identifies four common causes of vitamin use and birth defects, all of them potential confounders: pre-natal care, socioeconomic status (SES), difficulty conceiving, and maternal genetics. Vitamin use, birth defects, and the potential confounders are represented as *nodes* of the DAG, while the causal relationships between them are represented as arrows, or *directed edges*. The DAG is *acyclic* in the sense that no ordered sequence of arrows or directed edges leads back to the node from which the sequence began.

The DAG in Fig. 10.2 encodes several causal assumptions:

- Pre-natal care affects both vitamin use and risk of birth defects.
- A history of difficulty conceiving affects the likelihood that expectant mothers seek pre-natal care.
- Maternal genetics is a common cause of difficulty conceiving and birth defects.
- SES affects access to pre-natal care as well as vitamin use.

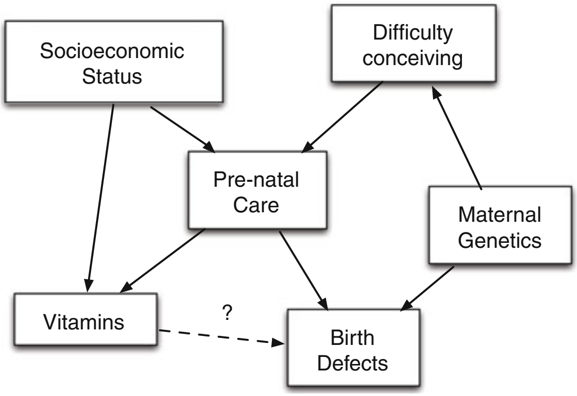

Fig. 10.2 Initial DAG for examining effects of vitamin use on birth defects

The preceding discussion of predictor selection for the second inferential goal suggests that we might want to control for all four hypothesized confounders of vitamin use. But do we really need to control for all of them? Not having to ascertain maternal genetics would save money and increase study participation, and a smaller model would likely be more efficient statistically.

10.2.5.2 Backdoor Paths

The DAG in Fig. 10.2 can be used to identify a minimum set of covariates we need to control for. To do this, we need to examine *backdoor paths* between vitamin use and birth defects. *Paths* are sequences of edges connecting two nodes, without regard to their direction. There are a total of five distinct paths connecting vitamin use and birth defects. Only one of these begins with a directed edge *from* vitamin use, specifically the path leading directly to birth defects, representing the hypothesized causal effect of interest; this is not a backdoor path. The other four paths connecting vitamin use and birth defects are backdoor paths, because they all include a directed edge leading *to* vitamin use:

(1) Vitamin use ← pre-natal care → birth defects
(2) Vitamin use ← SES → pre-natal care → birth defects
(3) Vitamin use ← pre-natal care ← difficulty conceiving ← maternal genetics → birth defects
(4) Vitamin use ← SES → pre-natal care ← difficulty conceiving ← maternal genetics → birth defects

Note that this DAG includes no paths beginning with a directed edge from vitamin use and passing through one or more nodes on the way to birth defects. Such indirect paths via mediators would not be considered backdoor paths.

10.2.5.3 Colliders

Pre-natal care is a so-called *collider* on the fourth backdoor path between vitamin use and birth defects, because it is the *common effect* of SES and difficulty conceiving. Note that pre-natal care is *not* a collider on any of the other three backdoor paths; likewise none of the other covariates are colliders on any of the four backdoor paths. Rules for determining what we need to control for treat colliders differently from other covariates along backdoor paths between exposure and outcome.

10.2.5.4 Blocking Backdoor Paths

Backdoor paths between exposure and outcome may be *blocked* or remain *open*. If any of the four backdoor paths between vitamin use and birth defects remains open, we would expect to find a statistical association between them, even if there was no causal relationship; essentially, this is uncontrolled confounding. But if all the backdoor paths are blocked, then we would only expect a statistical association between vitamin use and birth defects if a causal relationship links them. Whether any of the four backdoor paths remain open depends on whether it includes a collider, and what we control for in the statistical model we use to estimate the effect of vitamin use on birth defects. Specifically,

(1) A backdoor path is blocked, provided we control for at least one noncollider on the path. Thus, we can efficiently block the first three backdoor paths between vitamin use and birth defects by controlling for pre-natal care, because it is a noncollider on all those paths.
(2) A backdoor path including a collider is blocked, provided we do *not* control for the collider in the statistical model. Controlling for a collider induces a negative correlation between its common causes, opening an additional backdoor path, as shown in Fig. 10.3. To block this path, the model must control for a noncollider on the newly opened path.

Thus, the DAGs in Figs. 10.2 and 10.3 imply that we could obtain an unbiased estimate of the causal effect of vitamin use on birth defects using a statistical model in which we parsimoniously controlled for pre-natal care as well as one of the other three hypothesized confounders: SES, difficulty conceiving, or maternal genetics.

This pattern of confounding relationships, examined in a slightly simpler form by Greenland et al. (1999), illustrates that controlling for one apparently sufficient confounder may not be enough, if it is also a collider. Nonetheless, the solution is simple: controlling for just one additional factor will block the new backdoor path opened by controlling for the collider. Thus, the insight gained from the DAG might still make it possible to increase the efficiency of our study, relative to the more inclusive model selection strategy discussed earlier in this section.

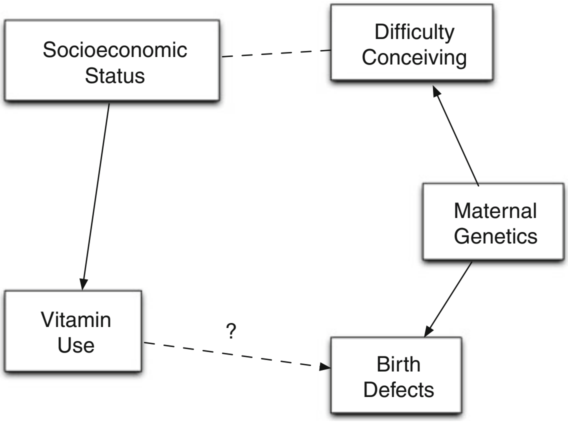

Fig. 10.3 Additional backdoor path opened by controlling for pre-natal care

10.2.5.5 Vulnerability to Assumptions

This result may be vulnerable to several assumptions implicit in the DAG in Fig. 10.2. Specifically, we may question whether

- *SES affects birth defects only through its effects on pre-natal care and vitamin use.* An additional pathway may result from environmental exposures, which are concentrated among the poor and minorities.
- *Difficulty conceiving affects vitamin use only through uptake of pre-natal care.* An additional pathway could be opened by the huge market for over-the-counter dietary supplements.
- *There is no direct link between maternal genetics and SES.* So-called population stratification suggests that the prevalence of genetic factors causing birth defects may differ by race/ethnicity. This opens a complicated causal pathway from maternal genetics to SES, mediated by racial and class discrimination.

If these concerns are valid, then there are three additional backdoor paths we might need to block, as shown in Fig. 10.4:

(1) Vitamin use ← SES → birth defects
(2) Vitamin use ← difficulty conceiving ← maternal genetics → birth defects
(3) Vitamin use ← SES ← maternal genetics → birth defects

Thus, we would need to control for pre-natal care and SES, as well as either difficulty conceiving or maternal genetics.

10.2.5.6 Colliders We Should Not Adjust For

DAGs can also help us avoid adjusting in cases where this will *induce* bias. For example, suppose we hypothesized the causal relationships in Fig. 10.5. In this

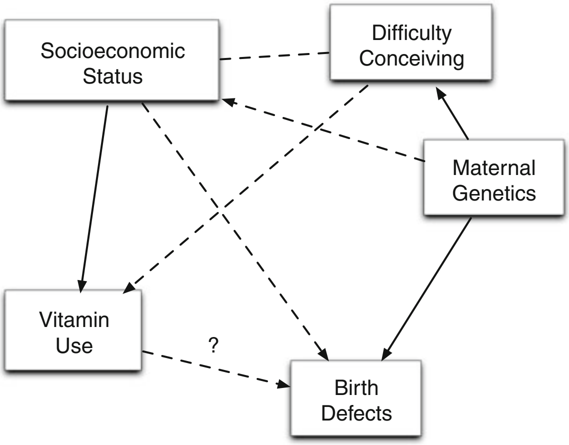

Fig. 10.4 Plausible additional causal pathways affecting birth defects

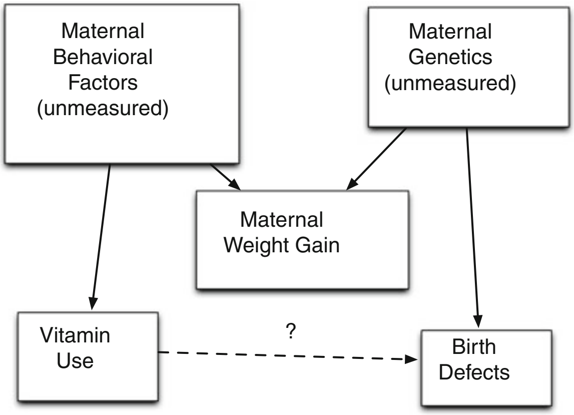
Maternal Behavioral Factors (unmeasured)
Maternal Genetics (unmeasured)
Maternal Weight Gain
Vitamin Use
?
Birth Defects

Fig. 10.5 Collider on pathway with unmeasured confounders

DAG, maternal weight gain is not a confounder of vitamin use, and so does not need to be adjusted for. As in Fig. 10.2, this depends on the *absence* of directed edges, in this case from maternal weight gain to vitamin use and birth defects. Their absence is based on substantive arguments: specifically, that most birth defects are caused by genetics and/or toxic exposures, with no prior evidence for an independent effect of weight gain; and that the perceptions of vitamin efficacy and inadequate diet, not maternal weight gain, are the primary motivations for vitamin use.

However, maternal weight gain *is* a collider on a backdoor path involving maternal behavioral and genetic factors, both unmeasured. Adjusting for maternal weight gain would *induce* bias in this case, by opening the backdoor path; moreover, we would be unable to block this path by adjusting for either of the two noncolliders, because they are unmeasured. Assuming the DAG in Fig. 10.5 is correct, it could prevent us from making this error, on the mistaken principle of adjusting for *any* possible confounder, without more carefully considering causal relationships.

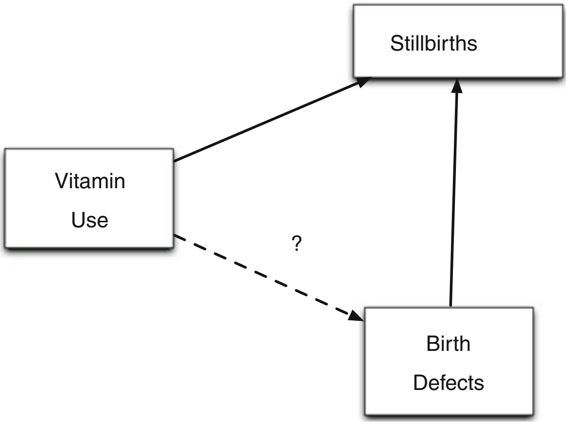

Fig. 10.6 Common cause of exposure and outcome is a collider

Similarly, Fig. 10.6 shows stillbirths, potentially reduced by vitamin use and increased by birth defects, as a common *effect* of exposure and outcome; in contrast, a confounder is a common cause. As a collider on the backdoor path between vitamin use and birth defects, stillbirths should not be adjusted for.

10.2.5.7 More About DAGs

The case of vitamin use and birth defects shows that using DAGs to identify a minimum set of covariates that must be controlled for may rest on strong assumptions that certain directed edges are *absent* from the DAG, assumptions that may be easy to second-guess, especially in newer fields of research. In that case, the safe course is to include the additional directed edges in the DAG and make sure that the resulting backdoor paths are blocked, either by a collider that is not controlled for in the statistical model, or by a noncollider that is.

At the same time, backdoor paths of the kind shown if Fig. 10.5 should be interpreted with caution if evidence for the unmeasured factors is unconvincing, or their effects are thought to be weak. In this case, leaving the backdoor path unblocked may not induce substantial bias. Greenland (2003) shows that bias from controlling for the common effects of exposure and outcome, as in Fig. 10.6, may often be comparable in magnitude with bias from not controlling for a common cause of exposure and disease. In contrast, biases from controlling for a collider as shown in Fig. 10.5 may be smaller.

DAGs are also useful for determining whether to adjust for the baseline outcome in analyses of pre-post change scores, as discussed in Sect. 7.3.1. Glymour et al. (2005) use DAGs to show that if exposure affects outcome *levels* at baseline (regardless of whether it affects subsequent changes), and the baseline outcome is measured with error, then bias results from adjusting for baseline. Similarly, so-called *horse-racing bias* arises if changes have already begun at baseline, and unmeasured causes of change affect both the baseline and follow-up outcomes. In both cases, the baseline outcome is a collider on a backdoor path from the primary predictor to the observed change. Since the common cause of the baseline outcome and change is by definition unmeasured, the resulting bias cannot be removed by adjustment.

In contrast, it *is* legitimate to adjust for the baseline outcome in estimating the effect of treatment on pre-post changes in a randomized trial, even though both outcomes are measured with error (Crager 1987). In this case, the directed edge from treatment to the baseline outcome is absent, so there is no backdoor path from treatment to change, with the corollary that the baseline outcome is not a collider.

In addition, Herńan et al. (2004) show how DAGs can be used to analyze the potential for selection bias. In particular, they show that restricting study entry according to participant characteristics is equivalent to adjusting for a collider, if common causes link the qualifying characteristics to both exposure and outcome. This approach also explains why informative censoring or dropout in longitudinal studies can induce bias. In contrast to the biases analyzed by Glymour et al. (2005), these biases can potentially be avoided by measuring and adjusting for the common causes linking exposure, outcome, and selection.

In summary, DAGs are a useful tool for thinking through what we need to adjust for in analyses focusing on the effect of a primary predictor, as well as what needs to be omitted, at least at the initial stages of an analysis. At the same time, overcomplicated DAGs should not stop progress—small biases from residual confounding or collider bias may not result in qualitatively mistaken inferences.

10.2.6 Randomized Experiments

In clinical trials and other randomized experiments, the intervention is the predictor of primary interest. Other predictors are, in expectation, uncorrelated with the intervention, by virtue of randomization. Thus, in the regression model used to analyze an experiment, covariates do not usually need to be included to rule out confounding of assignment to the intervention. However, there are several other reasons for including covariates in the models used to analyze experiments.

- *Making valid inferences in stratified designs.* Design variables in stratified designs need to be included to obtain correct standard errors, CIs, and P-values. At issue is the potential for clustering of outcomes within strata, potentially violating the assumption of independence (Chap. 7). Thus, analyses of multicenter clinical trials now commonly take account of clinical center, even though random and equal allocation to treatment within center ensures that treatment is in expectation uncorrelated with this factor. Clustering within center can arise from differences in the populations studied and in the implementation of the intervention.
- *Increasing precision and power in experiments with continuous outcomes.* Adjusting for important baseline predictors of a continuous outcome can increase

the precision of the treatment effect estimate by reducing the residual error; because the covariates are in expectation uncorrelated with treatment, the variance inflation factor described in Sect. 4.2.2 is usually negligible. However, Beach and Meier (1989) use simulations to suggest that adjustment may on average increase squared error of the treatment effect estimate in smaller studies or when the selected covariates are not strongly predictive of the outcome. They also explore the difficulties in selecting a reasonable subset of the many baseline covariates typically measured, and conclude that adjusting for covariates which are both imbalanced and strongly predictive of the outcome has the largest expected effect on the statistical significance of the treatment effect estimate. We support adjustment for important prognostic covariates in trials with continuous endpoints, but also endorse the stipulation of Hauck et al. (1998) that the adjusted model should be pre-specified in the study protocol, to prevent *post hoc* "shopping" for the set of covariates which gives the smallest treatment effect P-value.

- *"De-attenuating" the treatment effect estimate and increasing power in experiments with binary or failure time outcomes.* In contrast to linear models for continuous outcomes, omission of important but balanced predictors, including the stratification variables mentioned previously, from a logistic (Neuhaus and Jewell 1993; Neuhaus 1998) or Cox model (Gail et al. 1984; Schmoor and Schumacher 1997; Henderson and Oman 1999) used to analyze binary or failure time outcomes attenuates the treatment effect estimate. Hypothesis tests remain valid when the null hypothesis holds (Gail et al. 1988), but power is lost in proportion to the importance of the omitted covariates (Lagakos and Schoenfeld 1984; Begg and Lagakos 1993). Note, however, that adjustment for *im*balanced covariates can potentially move the treatment effect estimate *away* from as well as toward the null value, and can decrease both precision and power. In their review, Hauck et al. (1998) recommend adjustment for influential covariates in trials analyzed using logistic and Cox models. Their rationale is not only increased efficiency, but also that the adjusted or de-attenuated treatment effect estimates are more nearly interpretable as *subject specific*—in contrast to *population averaged*, a distinction that we explain in Sect. 7.5. We cautiously endorse adjustment for important covariates in trials with binary and failure time endpoints, but only if the adjusted model can be pre-specified and adjustment is likely to make the results more, not less convincing to the intended audience.
- *Adjusting for baseline imbalances.* Adjusted analyses are often conducted when there are apparent imbalances between groups, which can arise by chance, especially in small studies, or because of problems in implementing the randomization. The treatment effect estimate can be badly biased when strongly predictive covariates are imbalanced, even if the imbalance is not statistically significant. It is of course not possible to pre-specify such covariates, but adjustment is commonly undertaken in secondary analyses to demonstrate that the inferences about the treatment effect are not qualitatively affected by any apparent baseline imbalance. Note that the precision and statistical significance of the treatment effect estimate can be eroded by adjustment in this case, whether the endpoint

is continuous, binary, or a failure time. However, a difficult problem can arise when the selection of covariates to adjust for makes a substantive difference in interpretation, as Beach and Meier (1989) show in a re-analysis of time-to-event data from the Chicago Breast Cancer Surgery Study (Meier et al. 1985). In this small trial ($n = 112$), where the unadjusted treatment effect estimate just misses statistical significance ($P = 0.1$), different sets of covariates give qualitatively different results, with some adjusted models showing a statistically significant treatment effect and others weakening and even reversing the direction of the estimate.

10.3 Identifying Multiple Important Predictors

When the focus is on evaluating a predictor of primary interest, covariates are included in order to obtain a minimally confounded estimate of the association of the main predictor with the outcome. A good model rules out confounding of that association as persuasively as possible. However, broadening the focus to multiple important predictors of an outcome can make selecting a single best model considerably more difficult.

For example, inferences about most or all of the predictors retained in the model are now of primary interest, so overfitting and false-positive results are more problematic, particularly for novel associations not strongly motivated a priori. Effect modification or interaction will usually be of interest, but systematically assessing the large number of possible interactions can easily lead to false-positive findings, some at least not easily rejected as implausible. It may also be difficult to choose between alternative models that each include one variable from a collinear pair or set. Mediation is also more difficult to handle, to the extent that the overall effect of any predictor as well as its direct and indirect effects may be of interest. In this case, multiple, nested models may be required, as outlined in Sect. 4.4. Especially in the earlier stages of research, modeling these complex relationships is difficult, prone to error, and likely to be an iterative process. In some cases, a series of models, possibly including interactions, might be necessary to give a full and interpretable picture.

10.3.1 Ruling Out Confounding Is Still Central

In exploratory analyses to identify the important predictors of an outcome, confounding remains a primary concern—in this case, for any of the independent predictors of interest. Thus, some of the same strategies useful when a single predictor is of primary interest are likely to be useful here. In particular, relatively large models, including variables thought necessary for face validity, are preferable. Ideally, the model can be specified a priori. However, as in the previous section, small sample size and high correlation between predictors may limit the number of variables that can be included. In this case, we recommend using backward selection with a liberal retention criterion. We discuss these issues in more detail in Sects. 10.4.1 and 10.4.2.

Simplifying the problem by treating each of the candidate predictors in turn as a predictor of primary interest, using the procedures from the previous section, is not a particularly satisfactory solution in our view. This can result in as many different models as there are predictors of interest, especially if covariates are retained because removing them changes the coefficient of the predictor of interest. Such a description of the data is uneconomical and hard to reconcile with an internally consistent causal model. Furthermore, missing values can result in the different models being fit to different subsets of the data.

10.3.2 Cautious Interpretation Is Also Key

What principally differs in this context is that *any* of the associations in the final model may require substantive interpretation, not just the association with a primary predictor. This may justify a more conservative approach to some minor aspects of the model; for example, poorly motivated and implausible interactions might more readily be excluded. In addition, well-motivated choices among any set of highly correlated predictors would need to be made.

However, we do not recommend "parsimonious" models that only include predictors that are statistically significant at $P < 0.05$ or even stricter criteria, especially with small samples, because the potential for residual confounding in such models is substantial. At the same time, we do not recommend explicit correction for multiple comparisons, since in an exploratory analysis it is far from clear how many comparisons to correct for, and by how much. This is in contrast to analyses evaluating multiple outcomes of a single treatment, as discussed in Sect. 13.4.1, where adjustment is almost certainly needed.

A better approach is to interpret the results of a larger model cautiously, especially novel, implausible, weak, and borderline statistically significant associations, to report model selection procedures, including the complete list of covariates considered, and to be aware of the potential inflation of type-I error, listing this as a limitation in published descriptions.

A more radical alternative, briefly discussed in Sect. 10.6, is to use methods for developing prediction models, based on minimizing prediction error, often via cross-validation. For example, the LASSO, discussed in Sect. 10.1.4, drops the least important variables and shrinks the less precisely estimated coefficients for others that are retained. Some of these methods provide direct measures of so-called *variable importance*, the implicit focus of this inferential goal. Drawbacks often include the lack of P-values and CIs, and the difficulty of accounting for mediating relationships and retaining variables for face validity.

10.3.3 Example: Risk Factors for Coronary Heart Disease

Vittinghoff et al. (2003) used multipredictor Cox models to assess the associations between risk factors and CHD events among 2,763 postmenopausal women with established CHD. Because of the large number ($n = 361$) of outcome events, it was possible to include all previously identified risk factors that were statistically significant at $P < 0.2$ in unadjusted models and not judged redundant on substantive grounds in the final multipredictor model. Among the 11 risk factors judged to be important on both substantive and statistical grounds were six noted by history (nonwhite ethnicity, lack of exercise, treated diabetes, angina, congestive heart failure, ≥ 2 previous heart attacks) and five that were measured (high blood pressure, lipids including LDL, HDL, and Lp(a), and creatinine clearance).

For face validity and to rule out confounding, the final model also controlled for other known or suspected CHD risk factors, including age, smoking, alcohol use, and obesity, although these were not statistically significant in the adjusted analysis. Mediation of obesity and diabetes, both shown to be associated with risk in single-predictor models, was covered in the discussion section of the paper. The model also controlled for a wide range of CHD-related medications, but because these effects were not of direct interest and hard to interpret, estimates were not presented. However, interactions between risk factors and relevant treatments were examined, on the hypothesis that treatments might modify the association between observed risk factor levels and future CHD risk; the final model included interactions that were statistically significant at $P < 0.2$.

10.3.4 Allen–Cady Modified Backward Selection

Flexible predictor selection procedures, including conventional backward selection, are known to increase the probability of making at least one type-I error. A backward selection procedure (Allen and Cady 1982) based on a ranking of the candidate variables by importance can be used to help avoid false-positive results, while still reducing the number of covariates in the model. In this procedure, a set of variables may be forced into the model, including predictors of primary interest, as well as confounding variables thought important for face validity. The remaining candidate variables would then be ranked in order of importance. Starting with an initial model including all covariates in these two sets, variables in the second set would be deleted in order of ascending importance until the first variable meeting a criterion for retention is encountered. Then the selection procedure stops.

This procedure is special in that only the remaining variable hypothesized to be least important is eligible for removal at each step, whereas in conventional backward selection, any of the predictors not being forced into the model is eligible. False-positive results are less likely because there is only one pre-specified sequence of models, and selection stops when the first variable not meeting the criterion for removal is encountered. In contrast, conventional stepwise procedures and especially best subsets search over broader classes of models.

10.4 Some Details

10.4.1 Collinearity

In Sect. 4.2, we saw that the variance of the regression coefficient estimate for predictor x_j, increases with r_j, the multiple correlation between x_j and the other predictors in the model. When r_j is large, the estimate of β_j can become quite imprecise. Consider the case where two predictors are fairly highly correlated ($r \geq 0.80$). When both are included in the model, the precision of the estimated coefficient for each can be severely degraded, even when both variables are statistically significant predictors in simpler models that include one but not both. In the model including both, an F-test for the joint effect of both variables may be highly statistically significant, while the variable-specific t-tests are not. This pattern indicates that the two variables jointly provide important information for predicting the outcome, but that neither is necessary over and above the other. With modern computers, problems in estimating the independent effects of highly correlated predictors no longer arise from numeric inaccuracy in the computations. Rather, the information is coming from both variables jointly, which makes them both seem unimportant in t-tests evaluating their individual contributions.

> *Definition*: *Collinearity* denotes correlation between predictors high enough to degrade the precision of the regression coefficient estimates substantially for some or all of the correlated predictors.

How we deal with collinear predictors depends in part on our inferential goals. For a prediction model, inference on individual predictors is not of direct interest. Rather, if inclusion of collinear variables decreases prediction error, then it is legitimate to include them both. In this case, cross-validation of the target measure of PE can be used to decide which of a collinear set of predictors to include.

Alternatively, suppose that one of two collinear variables is a predictor of primary interest, and the other is a confounder that must be adjusted for on substantive grounds. If the predictor of interest remains statistically significant after adjustment, then the evidence for an independent effect is usually convincing. In small data sets especially, it would be necessary to demonstrate that the finding is not the result of a few influential points, and where the data do not precisely meet model assumptions, to show that the inferences are robust, possibly using the bootstrap methods introduced in Sect. 3.6. Alternatively, if the effects of the predictor of interest are clearly confounded by the adjustment variable, we would also have a clearcut result. However, in cases where neither is statistically significant after adjustment, we may need to admit that the data are inadequate to disentangle their effects.

In contrast, where the collinearity is between adjustment variables and does not involve the predictor of primary interest, then inclusion of the collinear variables can sometimes be justified. In this case, information about the underlying factor being adjusted for may be increased, but the precision of the estimate for the predictor of interest is unaffected. To see this, consider evaluating the effect of diabetes on HDL, adjusting for BMI. In Sect. 4.7, we found that a quadratic term in BMI added significantly to the model. However, `BMI` and its square are clearly collinear ($r = 0.99$). If instead we first "center" `BMI` (i.e., subtract off its sample mean before computing its square), the collinearity disappears ($r = 0.46$). However, the estimate for diabetes and its standard error are unchanged whether or not we center `BMI` before computing the quadratic term. In short, collinearity between adjustment variables is unlikely to matter.

Finally, when we are attempting to identify multiple independent predictors, an attractive solution is to choose on substantive grounds, such as plausibility as a causal factor. Otherwise, it may make sense to choose the predictor that is measured more accurately or has fewer missing values. As in the case of a predictor of primary interest, the multivariable model may sometimes provide a clear indication of relative importance, in that one of the collinear variables remains statistically significant after adjustment, while the others appear to be unimportant. In this case, the usual course would be to include the statistically significant variable and drop the others.

10.4.2 Number of Predictors

The rationale for inclusive predictor selection rules, whether we are assessing a predictor of primary interest or multiple important independent predictors, is to obtain minimally confounded estimates. However, this can make regression coefficient estimates less precise, especially for highly correlated predictors. At the extreme, model performance can be severely degraded by the inclusion of too many predictors.

Rules of thumb have been suggested for number of predictors that can be safely included as a function of sample size or number of events. A commonly used guideline prescribes at least ten observations for each predictor; with binary or survival outcomes the analogous guideline specifies ten events per predictor (Peduzzi et al. 1995, 1996; Concato et al. 1995). The rationale is to obtain adequately precise estimates, and in the case of the logistic and Cox models (Chaps. 5 and 6), to ensure that the models behave properly.

Such guidelines are useful as flags for potential problems, but need not be inflexibly applied. Their primary limitation is that the precision of coefficient estimates depends on other factors as well as the number of observations or events per predictor (Vittinghoff and McCulloch 2007). In particular, recall from Sect. 4.2 that the variance of an estimated regression coefficient in a linear model depends on the residual variance of the outcome, which is generally reduced by the inclusion of important covariates. Precision also depends on the multiple correlation between a predictor of interest and other variables in the model. Thus, addition of covariates that are at most weakly correlated with the primary predictor but explain substantial outcome variance can actually improve the precision of the estimate for the predictor

Table 10.3 Cox models for DVT-PE

Predictor	RH (95% Confidence interval)				*P*-values	
variable	11-Predictor model		5-Predictor models		Wald	LR
HT vs. placebo	2.7	(1.4–5.2)	2.7	(1.4–5.1)	0.002	0.001
≥ 53 at LMP	3.6	(2.0–6.4)	3.3	(1.8–5.8)	< 0.001	< 0.001
Inpatient surgery	4.3	(2.1–8.7)	4.7	(2.3–9.5)	< 0.001	< 0.001
Hospitalization	5.6	(2.9–11)	6.7	(3.6–13)	< 0.001	< 0.001
Hip fracture	5.9	(0.8–46)	6.6	(0.9–51)	0.09	0.18
Leg fracture	17.3	(5.1–58)	14.1	(4.2–47)	< 0.001	< 0.001
Cancer	4.1	(1.7–9.7)	3.5	(1.5–8.4)	0.002	0.006
Nonfatal MI	6.0	(2.3–16)	4.4	(1.7–11)	< 0.001	0.002
Stroke/TIA	0.9	(0.1–6.5)	0.9	(0.1–6.4)	0.88	0.88
Aspirin use	0.4	(0.2–0.7)	0.4	(0.2–0.6)	0.003	0.004
Statin use	0.4	(0.2–0.9)	0.4	(0.2–0.7)	0.02	0.02

of interest. In contrast, addition of just one collinear predictor can degrade its precision unacceptably. In addition, the allowable number of predictors depends on effect size, with larger effects being more robust to multiple adjustment than smaller ones.

Rather than applying such rules categorically, we recommend that problems potentially stemming from the number of predictors be assessed by checking for high levels of correlation between a predictor of interest and other covariates, and for large increases in the standard error of its estimated regression coefficient when additional variables are included. For logistic and Cox models, consistency between Wald and LR test results is another useful measure of whether there are enough events to support the number of predictors in the model. Additional validation of a relatively inclusive final model is provided if a more parsimonious model with fewer predictors gives consistent results, in particular for the predictor of interest. If problems do become apparent, a first step would be to make the criterion for retention in backward selection more conservative, possibly $P < 0.15$ or $P < 0.10$. It would also make sense to consider omitting variables included for face validity which do not appear to confound a predictor of primary interest.

An analysis of risk factors for deep-vein thrombosis and pulmonary embolism (DVT-PE) among postmenopausal women in the HERS cohort (Grady et al. 2000) is an example of stable results despite violation of the rule of thumb that the number of events per predictor should be at least 10. In this survival analysis of 47 DVT-PE events, 11 predictors were retained in the final model, so that there were only 4.3 events per predictor. However, the largest pairwise correlation between the selected risk factors was only 0.16 and most were below 0.02. As shown in Table 10.3, estimates from the 11-predictor model were consistent with those given by 5-predictor models, in accord with the rule of thumb, which omitted the less important predictors. Although CIs were wide for the strongest and least common risk factors, this was also true for the 5-predictor models. Finally, *P*-values for the Wald and LR tests based on the larger model were highly consistent.

10.4.3 Alternatives to Backward Selection

Some alternatives to backward selection include best subsets; sequential (so-called *greedy*) procedures, including forward and stepwise selection; and bivariate screening.

- *Best subsets* screens models including all possible subsets of the candidate predictors in a user-specified range of model sizes, using a summary measure such as adjusted R^2 to compare models. This computer-intensive procedure is implemented in SAS for some models, but not in Stata. It was also the underlying approach of the cross-validation screening described in Sect. 10.1.6, but did require prior simplification to reduce the computational burden.
- *Forward selection* begins with the null model with only the intercept, then adds variables sequentially, at each step adding the variable that promises to make the biggest additional contribution to the current model.
- *Stepwise* methods augment the forward procedure by allowing variables to be removed if they no longer meet an inclusion criterion after other variables have been added. Stata similarly augments backward selection by allowing variables to re-enter after removal. As compared to best subsets, these three sequential procedures are more vulnerable to missing good alternative models that happen not to lie on the sequential path. This implies that plausible alternatives to models selected by stepwise procedures should be examined.
- In *bivariate screening*, candidate predictors are evaluated one at a time in single-predictor models. In some cases, all predictors that meet the screening criterion are included in the final model; in other cases, screening is used as a first step to reduce the number of predictors then considered in a backward, forward, stepwise, or best subsets selection procedure. Orwoll et al. (1996) used a variant of this procedure, including all variables statistically significant at $P < 0.05$ in two-predictor models adjusting for age.

Note that only observations with complete data on all variables under consideration are used in automated selection procedures. The resulting subset can be substantially smaller than the data set used in the final model, and unrepresentative. When implemented by hand, different subsets are commonly used at different steps, for the same reason, and this can also affect results. Findings which depend on the inclusion or exclusion of subsets of observations should be carefully checked.

10.4.3.1 Why We Prefer Backward Selection

The principal advantage of backward selection is that negatively confounded sets of variables are less likely to be omitted from the model (Sun et al. 1999), since the complete set is included in the initial model. Best subsets shares this advantage. In contrast, forward and stepwise selection procedures will only include such sets if at least one member meets the inclusion criterion in the absence of the others.

Univariate screening will only include the complete set if all of them individually meet the screening criterion; moreover, this difficulty is made worse if a relatively conservative criterion is used to reduce the number of false-positive findings in an exploratory analysis.

A disadvantage of backward selection is that initial deletions may be badly determined if the list of candidate predictors is too large for the number of observations or events. In this case, bivariate screening with a liberal criterion can be used to eliminate the weakest predictors; in addition, the Stata stepwise procedure allowing variables to re-enter affords some protection against this problem. More generally, sensitivity analyses using forward and/or stepwise in addition to backward selection are useful for showing whether results are robust to the model selection procedure used

10.4.4 Model Selection and Checking

Section 4.7 focused on methods for checking the linear model which make use of the residuals from a multipredictor model rather than examining bivariate relationships. There, we took as a given that the predictors had already been selected. However, transformation of the outcome or of continuous predictors can affect the apparent importance of predictors. For example, in Sect. 4.6.4 we saw that the need for an interaction between treatment with HT and the baseline value of the outcome LDL was eliminated by analyzing treatment effects on percent rather absolute change from baseline. Alternatively, detection of important nonlinearities in the model checking step can uncover associations that were masked by an initial linear specification. As a consequence, predictor selection should be revisited after changes of this kind are made. And then, of course, the fit of the modified model would need to be rechecked.

10.4.5 Model Selection Complicates Inference

Underlying the CIs and P-values which play a central role in interpreting regression results is the assumption that the predictors to be included in the model were determined a priori without reference to the data at hand. In *confirmatory* analyses in well-developed areas of research, including phase-III clinical trials, prior determination of the model is feasible and important. In contrast, at earlier stages of research, data-driven predictor selection and checking are reasonable, often necessary, and certainly widely used. However, some of the issues raised for inference include the following.

- The chance of at least one type-I error can greatly exceed the nominal level used to test each term, leading to false-positive results with too-small P-values and too-narrow CIs.

- In small data sets, precision and power are often poor, so important predictors may well be omitted from the model, especially if a restrictive inclusion criterion is used. Conversely, in large data sets unimportant predictors are commonly included, reinforcing the need for cautious interpretation of novel, implausible, weak, and borderline statistically significant findings.
- Parameter estimates can be biased away from the null, owing to selection of estimates that are large by chance, sometimes called *testimation bias* (Steyerberg 2009). This bias is greater for relatively weak predictors.
- Choices between predictors can be poorly motivated, especially between collinear variables. Univariate screening provides no guidance for this problem. Moreover, predictor selection is potentially sensitive to addition or deletion of a few observations, especially when the predictors are highly correlated. Altman and Andersen (1989) propose bootstrap methods for assessing this sensitivity.

Predictor selection driven by P-values is subject to these pitfalls whether it is automated or implemented by hand. How seriously do these problems affect inference for our three inferential goals?

- *Prediction.* In many modern prediction methods, potentially large sets of candidate predictors are aggressively screened, but P-values are not used as the criterion. We implemented one such procedure in Sect. 10.1.6, and Breiman (2001) briefly reviews other modern methods which even more aggressively search over candidate models. However, use of GCV measures of prediction error as a criterion for predictor selection effectively protects against both overfitting and invalid inferences. In short, predictor selection does not adversely affect modern procedures for this inferential goal.
- *Evaluating a predictor of primary interest.* Iterative model checking and selection should likewise have relatively small effects on inference about a predictor of primary interest, since it is included by default in all candidate models. In fact, iterative checking and predictor selection should result in better control of confounding, a primary aim for this inferential goal. However, when the primary predictor is of borderline statistical significance, the issue of P-value shopping raised in Sect. 10.2.6 needs to be conscientiously handled, and sensitivity of results to predictor selection reported.
- *Identifying multiple important predictors.* Model selection most clearly complicates inference for this inferential goal, since CIs and P-values for any of the predictors are potentially of direct interest. Note that inclusion of variables for face validity, use of a loose inclusion criterion ($P < 0.2$), and the Allen–Cady procedure all reduce the potential impact of predictor selection on inference. Nonetheless, selection procedures should *only* be used with prior consideration of hypothesized relationships, careful examination of alternative models with other sets of predictors, checks on model fit and robustness, skeptical review of the findings for plausibility, and cautious interpretation of the results, especially novel, borderline statistically significant, and weak associations.

10.5 Summary

We have identified three inferential goals, and recommend predictor selection procedures appropriate to each of them.

For prediction, we recommend identifying candidate predictors and appropriate transformations well-supported by prior research. But in the common case where expert opinion and the literature do not provide sufficient guidance, we recommend exhaustive screening of candidate models to find the few models that minimize a generalized cross-validation measure of prediction error.

For evaluating a predictor of primary interest, we recommend using DAGs to specify hypothesized relationships between the primary predictor, potential confounders and mediators, and the outcome; caution should be used in eliminating variables based on any DAG that omits plausible but unestablished causal pathways. The selected model should include all generally accepted confounders required to ensure its face validity. Other potential confounders that turn out not to be important on statistical grounds can optionally be removed from the model using a backward selection procedure, but with a liberal inclusion criterion to minimize the potential for confounding. Especially in smaller data sets, care must be taken with the inclusion of covariates highly correlated with the predictor of interest, since these can unduly inflate the standard errors of the estimate of its effect. Negative findings for the primary predictor should be carefully interpreted in terms of the point estimate and CI, as described in Sect. 3.7.

For identifying multiple important predictors of an outcome, we recommend a procedure similar to that used for a single predictor of primary interest. A DAG mapping out hypothesized relationships between variables can be particularly useful. Strongly motivated covariates may be included by default to ensure the face validity of the model. The Allen–Cady modification of the backward selection procedure is useful for selecting from among the remaining candidate variables while limiting false-positive results. Negative, weak, and/or borderline statistically significant associations retained in the final model as much to control confounding of other associations as for their intrinsic plausibility and importance should be interpreted with particular caution.

10.6 Further Notes and References

Predictor selection is among the most controversial subjects covered in this book. Book-length treatments include Miller (1990) and Linhart and Zucchini (1986), while regression texts including Weisberg (1985) and Hosmer and Lemeshow (2000) address predictor selection issues at least briefly. The central place we ascribe to ruling out confounding in the second and third inferential goals owes much to Rothman and Greenland (1998), a standard reference in epidemiology that describes how substantive considerations can be brought to bear on predictor selection.

One promising method for ensuring adequate control of confounding is more or less exhaustive screening of candidate models with different covariate sets, some including interactions between covariates and/or restricted cubic splines for continuous confounders. As described in Sect. 10.1.4, these procedures use cross-validated prediction error as a model selection criterion to avoid overfitting, and avoid some pitfalls of P-value driven selection procedures, as discussed in Sect. 10.4.5. However, these methods can be difficult to implement, and are a focus of ongoing statistical research.

Both the theory and application of causal diagrams and models have been advanced substantially in recent years (Pearl 1995; Greenland et al. 1999) and give additional insights into situations where confounding can be ruled out a priori. However, these more advanced methods appear to be most useful in problems where causal pathways are more clearly understood than is our usual experience. Jewell (2004) and Greenland and Brumback (2002) explore the connections between causal diagrams, potential outcomes, and some model selection issues.

Chatfield (1995) reviews work on the influence of predictor selection on inference, while Buckland et al. (1997) propose using weighted averages of the results from alternative models as a way of incorporating the extra variability introduced by predictor selection in computing CIs. These would be particularly applicable to the second inferential goal of evaluating a predictor of central interest.

For a sobering view of the difficulty of validly modeling causal pathways using the procedures covered in this book and particularly this chapter, see Breiman (2001). From this point of view, computer-intensive methods validated strictly in terms of prediction error not only give better predictions but may also be more reliable guides to "variable importance"—another term for our third inferential goal of identifying important predictors, and with obvious implications for assessing a predictor of central interest.

10.7 Problems

Problem 10.1. Characterize the following contexts for predictor selection as prediction, evaluation of a primary predictor of interest, or identifying the important predictors of an outcome:

- examining the effect of treatment on a secondary endpoint in an RCT
- determining which newborns should be admitted to the neonatal intensive care unit (NICU)
- comparing a measure of treatment success between two surgical procedures for stress incontinence using data from a large longitudinal cohort study
- identifying risk factors for incident hantavirus infection.

Problem 10.2. Consulting Stata documentation, describe how the `sw:` command prefix with the `lockterm1`, `hier`, and `pr()` options can be used to implement the Allen–Cady procedure.

Problem 10.3. Think of an outcome under preliminary investigation in the area of your expertise. Following Allen and Cady's prescriptions, try to rank predictors of this outcome in order of importance. Are there any variables that you would include by default? Why?

Problem 10.4. Do any of the variables you have selected in the previous problem potentially mediate the effects of others in your list? If so, how would this affect your decision about what to include in the initial model? What series of models could you use to examine mediation? (See Sect. 4.5.)

Problem 10.5. Suppose you included an indicator for diabetes in a multivariable model estimating the independent effect of exercise on glucose. How would you interpret the estimate for exercise? Would you want to consider interactions between exercise and diabetes in this model? How would you deal with use of insulin and oral hypoglycemics?

Problem 10.6. Why are univariate screening and forward selection more likely to miss negatively confounded variables than backward deletion and best subsets?

Problem 10.7. Give an example of a "biologically plausible" relationship that has turned out to be false. Give an example of a biologically *im*plausible relationship that has turned out to be true.

Problem 10.8. Suppose you were using a logistic model to examine the association between a predictor and outcome of interest, and to rule out confounding you needed to include one or two more predictors than would be allowed by the rule of 10 events per variable. In comparing models with and without the two extra predictors, what might signal that you were asking the bigger model to do too much? How would the correlation between the extra variables and the predictor of interest influence your thinking?

10.8 Learning Objectives

(1) Describe and implement strategies for predictor selection for

- prediction
- evaluation of a primary predictor
- identifying multiple important predictors.

(2) Use a DAG to define hypothetical relationships among confounders, mediators, and the outcome.
(3) Be familiar with the drawbacks of predictor selection procedures.