persons whose costs exceed a specified threshold. In winsorizing, equal numbers of both high- and low-value cases are replaced by specified values: For example, the five lowest values are reset to their maximum value and the five highest to their minimum. Winsorizing aims to create more stable estimates of means for essentially symmetrically distributed distributions (Dixon and Tukey 1968; Dixon and Massey 1969). Some researchers use the terms winsorizing or truncation when referring to what we called topcoding.

Various studies involving risk adjustment have used topcoding. In deriving APACHE III models to predict ICU LOS, Knaus and colleagues (1993) capped all ICU LOSs at 40 days. The Society of Actuaries study of risk-adjustment models performed analyses three ways: including all costs at their original values; topcoding costs greater than $100,000 at $100,000; and topcoding costs greater than $50,000 at $50,000 (Cumming et al. 2002). R^2 values were higher on topcoded data (see Table 10.1). When the topcode threshold was $50,000, R^2 values were even larger than for analyses with topcoding at $100,000.

We prefer topcoding to trimming because the most expensive cases are often very important and should not be completely ignored. We note that predictions from an OLS model fit to topcoded data occasionally exceed the topcode threshold. Such "out-of-range" predictions rather reliably mark the very sickest people; there is nothing to be gained by trying to "fix" them.

The effect of trimming or topcoding outliers may vary depending on how the risk adjuster is constructed. Methods that generate only a few categories (e.g., that assign scores of 1 to 5) may have defined a highest-risk category with few cases that are likely to have the most extreme outcome values. Conceptually, such risk adjusters should perform relatively better when outliers are included in the analysis than when they are excluded. Risk-adjustment approaches employing a continuous scale may distinguish better among non-outlier cases but not perform as well when cases with extreme values are included.

How and whether to trim or topcode data are best decided within the context of specific studies. In a payment policy context, for example, topcoding cases greater than $100,000 is consistent with a payment system that covers all costs below this threshold but pays for more costly cases out of a separate pool. Trimming or topcoding always understates expected costs (i.e., produces systematically smaller predicted than actual outcomes). This can prove misleading in a payment system with no special mechanism for handling outlier costs.

Another common strategy for limiting the influence of high-value cases is to transform the dependent variable, generally using a logarithmic transformation; the value of the dependent variable is replaced by its natural logarithm, producing a distribution of outcome values that is closer to normal. Lognormal transformations fall within the Box-Cox family of transformations (Box, Hunter, and Hunter 1978; Atkinson 1985; Spitzer 1982). With their more normal distribution, transformed data more closely conform to the assump-

tions of OLS modeling. Transforming continuous outcome variables is most appropriate when the goal is to identify statistically significant predictors of the outcome (to better meet the assumptions underlying the statistical tests). However, when the goal is to predict the value of the dependent variable, such transformations are less useful.

Predictions based on a transformed outcome variable must be retransformed to the original scale before calculating statistical performance measures. In particular, comparing R^2 values from predicting logged versus actual outcomes is inappropriate. As discussed above, OLS maximizes R^2 for the data to which the model is fit—whether on an original or transformed scale. Regardless of which scale is used in model development, statistical measures of model performance should be computed in the original scale (e.g., dollars, days).

Duan's (1983) smearing estimator is a widely used, theoretically attractive approach for retransforming log-transformed data. This estimator is a number (the average of the retransformed residuals) by which each prediction is multiplied to correct for the known bias associated with the retransformation. However, retransformed predictions often fail to achieve as high an R^2 as a model fit directly to the untransformed data. This is no surprise, as the OLS algorithm specifically estimates parameters so that predictions minimize R^2 in the scale of interest. In our study of hospital-based severity measures, we evaluated R^2 both on untransformed LOS data and log-transformed data, using Duan's smearing estimator to retransform predictions before calculating R^2. For both pneumonia and hip fracture cases, R^2 values were as high or higher using models fit to LOS rather than log (LOS) (Iezzoni et al. 1996c; Shwartz et al. 1996a).

As an alternative to transforming the data, the generalized linear model (GLM) provides a comprehensive framework for developing multivariable models with nonnormally distributed data. GLM assumes that some function of the outcome, called a "link" function, can be modeled as a linear function of the predictors. GLM seeks parameters to predict outcomes in their natural scale directly rather than in a transformed scale. The GLM framework also allows for independent specification of how variances might vary as a function of the mean value of the outcome (e.g., costs for expensive cases typically vary more than costs for inexpensive ones). GLMs are described in a classic but sophisticated text by McCullagh and Nelder (1989), which assumes knowledge of mathematical likelihood. Several books describe GLMs, especially their implementation in various software or programming languages (Chambers and Hastie 1992; Ripley and Venable 1994).

Another alternative for limiting the influence of outliers is to seek models with the smallest mean absolute prediction error, also called the mean absolute deviation (MAD), rather than the smallest squared error. MAD is the average of the absolute value of $(Y - PRED)$. The "deviation" in MAD denotes the same quantity as "error" in the phrase "mean squared error" (MSE).

Although standard software packages [illegible] minimize MADs, analysts can nevertheless compare risk adjusters based on their MAD values. The Society of Actuaries took this approach in evaluating risk-adjustment models (Cumming et al. 2002). Table 10.4 shows R^2 values and MAD when the models were used prospectively with either no topcoding or with topcoding at $100,000. Values of MAD are smaller (demonstrating better average fit) with topcoded data than with actual costs. In general, models with higher R^2 values have lower MADs. However, relative model performance can differ depending on the statistical performance measures. For example, CDPS, DCGs, and ERGs all had the highest, and similar R^2 values. However, with MAD, ERGs appeared best and RxGroups second best.[7]

Although MAD may seem like a more natural measure of model fit than R^2, analysts are comfortable with standard deviation (SD) as a measure of the size of a typical error. In a given data set, the model whose residuals have the smallest SD is the model with the highest R^2.[8] Furthermore, models are most commonly used to predict group averages, making predictive validity hinge on minimizing the "average of n deviations." When predicting averages for moderately large groups, models that minimize MSE are preferable to those that minimize MAD.[9]

Finally, when considering variable transformations and nonlinear modeling, transparent modeling logic is very important. Developers likely need to convince nontechnical audiences about the merit of their approach (a kind of "face validity"). They also need to engage others who are particularly knowledgeable about the specific context of the work in critiquing the model (see Chapter 8). OLS modeling is relatively transparent. Lumley and collaborators (2002) used simulations to show that, with large data sets, OLS generally handles nonnormally distributed data well, despite its underlying assumptions about normality.

Interpreting R^2

With experience, analysts learn what values of statistical performance can be expected for "good" predictive models in specific contexts. For example, when predicting next year's health care costs for a Medicare enrollee from this year's data, actuaries and policymakers have looked favorably on models with R^2 values of 0.06 (Ash et al. 1989; Epstein and Cumella 1988). Although such small R^2 values indicate that individual *PRED*s might differ from the observed outcome by almost as much as if *PRED* always equaled the population mean,[10] these *PRED*s can identify subgroups with very different future costs.

The Society of Actuaries study demonstrates that validated R^2 values are now roughly 15 percent for models to predict next year's actual costs from demographic characteristics and this year's diagnoses or pharmacy claims. However, with more clinical information in administrative data sets (see Chapter 5), prospective models are beginning to produce R^2 values closer to 20 percent. Many observers remain unimpressed by models that explain only

TABLE 10.4 R^2 and Mean Absolute Deviation when Models to Predict Total Annual Cost Are Used Prospectively with Recalibrated Weights

Model	$R^2 \times 100$		MAD	
	Topcoded at $100,000	Full Range	Topcoded at $100,000	Full Range
ACG	14	10	2,100	2,193
CDPS	19	15	2,070	2,164
DCG	20	15	2,032	2,133
Medicaid Rx	17	12	2,062	2,159
RxGroups	19	13	2,014	2,113
RxRisk	15	11	2,091	2,187
ERG	20	15	1,983	2,079

SOURCE: Adapted from Tables 3.2 and 3.3, Cumming et al. (2002).

15 percent to 20 percent of the variation in cost. However, context is all important. Models that predict next year's actual costs will never approach 50 percent explanatory power because they cannot anticipate which specific individuals will experience acute illnesses or incur catastrophic expenses next year. Nonetheless, prospective models can identify the chronic problems that lead to large systematic differences in the future costs of groups.

When choosing among risk adjusters, analysts often look for the statistical performance measures reported in research publications. An important question is how well results from these published reports generalize to other settings and purposes. For instance, although APACHE II was developed explicitly to predict ICU deaths, it also has been used for non-ICU populations (Daley et al. 1988; Iezzoni et al. 1990; Keeler et al. 1990). APACHE III yielded a cross-validated R^2 of 0.15 to predict ICU LOS (Knaus et al. 1993). This figure comes from data drawn from a stratified random sample of 26 hospitals, plus 14 other volunteer hospitals that were primarily tertiary teaching facilities. The predominance of tertiary teaching hospitals raises questions about generalizability of this R^2 value to other hospital settings.

R^2 values depend on the cases in the database as well as the risk factors available for modeling. The dispersion of the independent and dependent variables significantly affects statistical performance, particularly R^2. Figure 10.1 shows three schematic diagrams of models to predict Y from a single variable X, where larger values for X indicate greater risks for poor outcomes. Graph A shows the classic bivariable normal situation. R^2 is approximately equal to:

$$1 - \frac{\text{variance in } (Y - PRED)}{\text{variance in } (Y)}.$$

The other two graphs in Figure 10.1 show what happens when only some of the data are available for modeling. In graph B, cases with extreme values of X are eliminated, producing less variation in Y but no change in

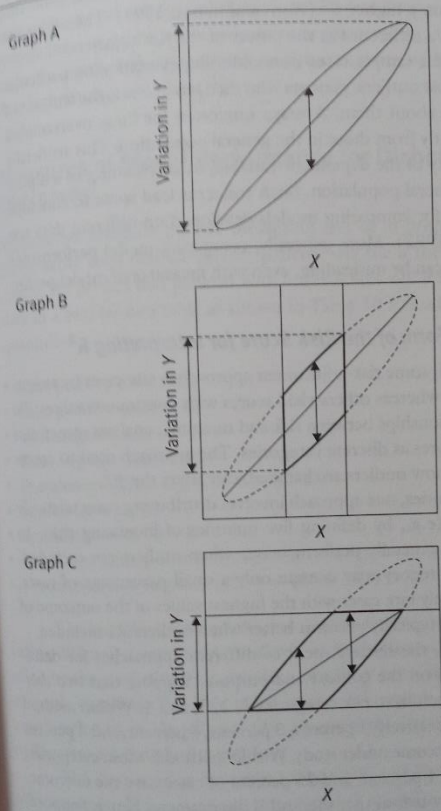

FIGURE 10.1 How Differences in the Data Modeled Affect R^2

NOTE: Graph A shows a regression line fit to a schematic scatterplot of bivariate normal data. In graph B, cases with extreme values of X have been removed; in graph C, cases with extreme values of Y have been removed. The shorter, vertical arrows in the body of each graph indicate variation in $(Y - PRED)$. The smaller the ratio of variation in $(Y - PRED)$ to variation in Y, the larger is R^2.

the variation in $(Y - PRED)$ for the remaining cases. In this situation, R^2 should decrease. In graph C, cases with extreme values of Y are removed, which reduces both kinds of variation. The net effect on R^2 is unpredictable.

Another example illustrates how characteristics of the data affect statistical performance. Suppose two risk adjustment methods are applied to two different data sets and are equally accurate in predicting outcomes for cases across the range of independent variables (i.e., SSE/n is identical for each risk

adjuster). The data set with more variability in the outcome variable generates a higher SST and, thus, a higher R^2 (Korn and Simon 1991). The difference in R^2 falsely suggests a difference in the power of the risk adjusters.

Certain data sets sample cases nonrandomly, oversampling particular types of cases (e.g., cost outliers, patients who died) to increase the amount of information available about them. Average outcomes for these oversampled cases often differ greatly from those in the general population. This artificially increases the variability of the dependent variable, usually leading to a higher R^2 than one for a general population. Such concerns lead some to conclude that R^2 is unsuitable for comparing models developed on different data sets (Cox and Wermuth 1992). More generally, comparing model performance on different data sets can be misleading, even with measures of model performance other than R^2.

Importance of the Form of the Risk Score for Interpreting R^2

As shown in Table 2.5, some risk-adjustment approaches rate cases by assigning categories of risk, whereas others yield scores with continuous values. To capture complex relationships between risk and outcome, analysts sometimes recode continuous scores as discrete categories. The approach used to create categories, including how outliers are handled, can affect the R^2.

To create categories, one approach involves distributing cases relatively evenly across groups (e.g., by defining five quintiles of increasing risk). In this instance, models generally perform better when outliers are excluded. In contrast, the top category may contain only a small percentage of cases, capturing those relatively rare cases with the highest values of the outcome of interest. These models typically perform better when outliers are included.

To some extent, the relative merit of different approaches for defining categories depends on the context and purpose. Suppose that two risk-adjustment methods each have risk categories 1, 2, 3, and 4. Within method A's four categories, respectively, 1 percent, 3 percent, 4 percent, and 8 percent of the cases have the outcome under study. Within method B's four categories, 0.5 percent, 3 percent, 4 percent, and 25 percent of cases have the outcome. At first glance, it seems obvious that method B discriminates better. Suppose, however, that 25 percent of cases fell into each of method A's four categories, whereas only 0.5 percent of cases were in each of method B's lowest and highest risk groups (categories 1 and 4). For many purposes, method A is more useful because it can meaningfully discriminate across relatively large numbers of people. If the goal is to find very small groups at particularly low or high risk, however, method B is better.

In comparing statistical performance measures across different approaches for creating discrete categories from continuous scales, it is important to know how the categories were defined and how cases distribute across the categories. If the categories were defined to maximize model performance in a development data set, the model's performance could deteriorate when