Therefore, other things being equal, the Republican Party starts off when it is the incumbent party with a predicted vote share that is larger than the predicted vote share for the Democratic Party when it is the incumbent party (50.7 versus 45.5). If the party variable is dropped and the coefficients are reestimated, the intercept is 47.5. This implies that if all the other variables have a value of zero, the Democratic Party starts off when it is the incumbent party with a predicted vote share of 47.5 and the Republican Party starts off when it is the incumbent party with a predicted vote share of 100 minus this, or 52.5. Either with or without the party variable included, the results show a bias in favor of the Republicans, other things being equal.

The coefficient for the war variable is only relevant for three elections. Its *t*-statistic is 1.99. The war variable is thus significant if we use a cutoff of 2.0 and round 1.99 up to 2.0.

Overall, the results seem very good. The errors are small except for 1992, and all the variables are significant except for the party variable. In particular, the *t*-statistics for the economic variables (the growth rate, inflation, and good news quarters) suggest that the economy does have an effect on the vote share; it is very unlikely that we would get these *t*-statistics if the economy did not affect the vote share. We cannot, however, relax because of the possible pitfalls lurking in the background. To these we now turn.

Possible Pitfalls

The main pitfall that we need to worry about is the possibility of data mining. Many variables were tried in arriving at the final results, and we have only 24 observations. It may be that by chance we have fit the data well, but in fact, the vote share is determined by other things. The following are the main things that were tried that may be subject to the data mining problem.

- Increments other than 0.25 were tried for the duration variable, and 0.25 was chosen because it gave the best results in terms of fit.
- Values other than 3.2 percent were tried for the cutoff for good news quarters, and 3.2 was chosen because it gave the best results in terms of fit.

- The particular treatment for the wars for the three elections was done because this led to an improved fit.
- Different periods for the growth rate were tried, and the particular one chosen, the first three quarters of the election year, gave the best results in terms of fit.
- Different periods for inflation were also tried, and the particular one chosen, the entire four-year period except for the last quarter, gave the best results in terms of fit.
- After the large error was made in 1992, an attempt was made to find reasons for it. This effort led to the choice of the good news quarters variable, which prior to 1992 had not been thought of. The good news quarters variable helps make the error for 1992 smaller because, as you can see from Table 1-1, there were only two good news quarters for the George Bush administration. President Bush is still predicted to win in 1992 in Table 3-1, but by less than he would be predicted to if it were not for the good news quarters variable.

With only 24 elections and all this searching, it is certainly possible that the results in Box 3-2 are a fluke and are not really right. As discussed in Chapter 2, one way of examining the seriousness of the data mining problem is to see how well future observations are predicted. If the results are a fluke, future predictions should not in general be very accurate. In particular, if the results are a fluke, the prediction for the 2012 election is not likely to be accurate, since no information about this election was used in getting the results. The prediction of the 2012 election is discussed later in this chapter.

An alternative approach to examining the data mining problem is to use only part of the observations to get the coefficients and then see how well these coefficients do in predicting the other observations. This is not as good a check as waiting because we have used information in the whole sample (both parts) to decide which variables to include, but at least the coefficients are obtained using only the information in the first part of the observations.

To perform this check, the best fitting set of coefficients was obtained using only the elections through 1960. In other words, the best fit was obtained for the 1916–1960 period (12 elections), and no data from