ONE

# *Time Series Analysis in the Social Sciences*

IN THE SOCIAL SCIENCES, data are usually collected across space, that is, across countries, cities, and so on. Sometimes, however, data are collected across time through repeated regular temporal observations on a single unit of analysis. With the data that are collected in this way and that are entered in chronological order, we explore the history of a variable to identify and explain temporal patterns or regularities in the variable. We also explore the relationships of a variable with other variables to identify the causes of the temporal pattern of the variable.

Time series analysis is not as frequently employed in the social sciences as regression analysis of cross-sectional data. However, this is not because time series analysis is less useful than regression analysis but because time series data are less common than cross-sectional data. It is the characteristics of the data at hand, not the usefulness of statistical techniques, which we consider to select between time series analysis and regression analysis.

When we deal with time series data, time series analysis can be more useful than ordinary least squares (OLS) regression analysis. Employing OLS regression analysis, we cannot appropriately model a time series, specifically its systematic fluctuations, such as seasonality and systematically patterned residuals (see chapter 2). As a result, the standard errors of regression coefficients are likely to be biased, and independent variables may appear to be statistically more significant or less significant than they actually are.

Time series analysis can be employed in several ways in the social sciences. The most basic application is the visual inspection of a long-term behavior (trend) of a time series (see chapter 2). For example, in order to survey the extent of partisan change in the southern region of the United States, Stanley (1988) visually inspected the percentages of Republicans, Democrats, and

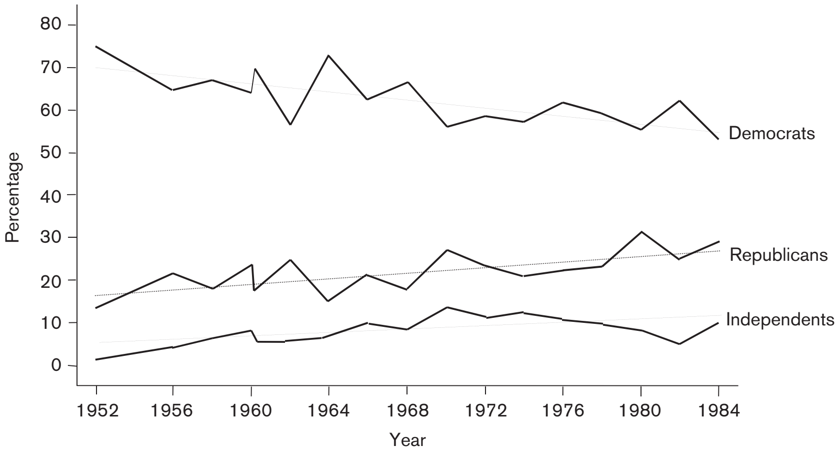


FIGURE 1. Party identification in the South, 1952–1984.
SOURCE: Stanley (1988), Figure 1, p. 65. Reproduced with permission of the University of Chicago Press.

independents from 1952 to 1984. As can be seen in figure 1, visual inspection is enough to show that both realignment and dealignment characterized southern partisan changes and that it was the Democratic Party that suffered the most from the change.

When we estimate trends, time series analysis is bivariate OLS regression analysis where the independent variable is Time with a regular interval. This time series analysis is called *univariate time series analysis* (see chapter 2). The trend in time series analysis is the slope of Time in bivariate OLS regression analysis. For example, Cox and McCubbins (1991) regressed the percentage of times individual legislators voted with their party leaders from the 73rd to the 96th Congress on Time. They showed that party voting significantly declined only for the Republicans (figure 2).

In many cases, a time series contains systematic short-term fluctuations other than a long-term trend. That is, observed values increase for a certain period and decrease for another period, rather than randomly fluctuating over the fitted linear line. These systematic patterns in time series variables should be removed to examine accurately the relationship between them. When systematic patterns are present in two time series variables, the correlation between the two can simply be a product of the systematic patterns (see chapters 2 and 6).

For example, Norpoth and Yantek's (1983) study of the lagged effect of economic conditions on presidential popularity raised a question about Mueller

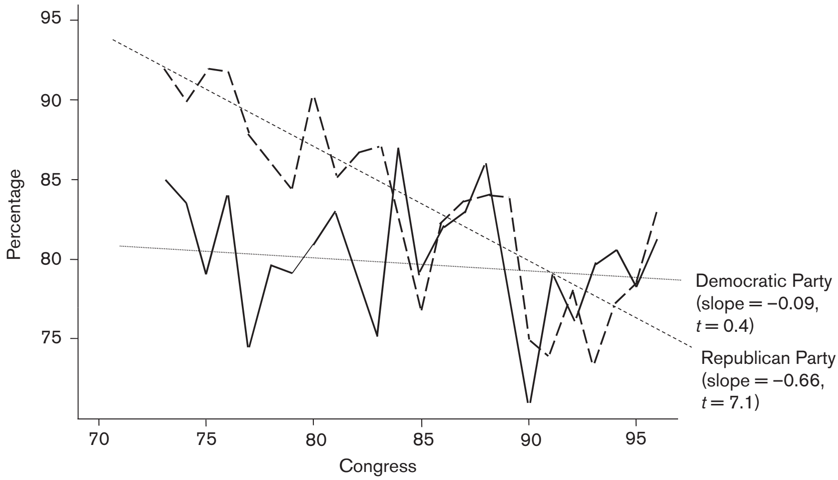


FIGURE 2. Average leadership support scores on the party agenda, 73rd–96th Congress.
SOURCE: Adapted from Cox and McCubbins (1991), Figures 1 and 2, pp. 557–558. Reproduced with permission of John Wiley & Sons Inc.

(1970, 1973), Kramer (1971), and Kernell (1978). Their estimates of economic effects, according to Norpoth and Yantek, are vulnerable to serial correlation within the independent variables, the monthly observations of unemployment or inflation. Norpoth and Yantek identified stochastic processes (ARIMA, explained in chapter 2) for the inflation series and for the unemployment series. They removed the estimated stochastic processes from the observed series. Since the inflation series and the unemployment series were no longer serially correlated, the relationship between inflation (or unemployment) and presidential popularity could not be an artifact of autocorrelation of the inflation series or of the unemployment series.[1] Norpoth and Yantek found that past values of the inflation and unemployment series did not significantly influence current approval ratings of presidents with any particular lag structure. This finding is not in accord with the conventional wisdom that the economy matters for presidential popularity and also for presidential electoral outcomes. Studies of presidential elections (Key 1966; Lewis-Beck 1988; Lockerbie 1992) present evidence that the national economy and the evaluation of a president's handling of the nation's economy do matter for the public support for the president. Norpoth and Yantek are reluctant to conclude that inflation and unemployment do not influence presidential popularity. They discuss some problems raised by the removal of the estimated stochastic processes from the observed series. Nonetheless, Norpoth and Yantek show that it is possible to lead to a very different finding when we ignore serial correlation in a time series versus when we remove it from the series.

Autocorrelation among residuals is a serious violation of a vital assumption concerning the error term in regression analysis (Achen 1982; Berry and Feldman 1985; Lewis-Beck 1980), and it is very likely to be present when we collect observations across time. With the serially correlated residuals, the least-squares estimates are still unbiased but may not be the best, with the minimum variance. Also, the significance tests and the confidence intervals for regression coefficients may be invalid. With time series analysis, we can directly check and estimate a systematic pattern that remains after we fitted a trend line to a time series (see chapter 3). If we are concerned only about the trend estimation, we can remove the systematic pattern from residuals before we fit a trend line to a time series by smoothing the time series (see chapter 5). In multiple time series analysis, we can deal with autocorrelated residuals in several different ways (see chapter 6). For example, we can estimate and then eliminate systematic patterns from each time series before we conduct multiple time series analysis. Alternatively, we can estimate a multiple regression model with autoregressive processes by adjusting regression coefficients according to estimated autocorrelation coefficients.

Once we estimate an autoregressive process of a time series, we can utilize the autoregressive process to determine how long the time series's time-dependent behavior or its impact on the dependent variable will persist (see chapter 6). For example, comparing the autoregressive parameters of two independent variables, Mackuen, Erikson, and Stimson (1989) show that the impact of consumer sentiment on aggregate-level party identification lasts longer than that of presidential approval, although the immediate impact of the former is smaller than that of the latter. As a result, the total impact of consumer sentiment on party identification is greater than that of presidential approval.

Time series analysis has perhaps been employed most frequently to forecast future outcomes, for example of presidential elections (see e.g. Lockerbie 2004, 2008; Norpoth and Yantek 1983; Norpoth 1995; Rosenstone 1983). In interrupted time series analysis, forecasted values are used as *counterfactuals* that represent a time series that we would have observed had there not been an intervention, such as the implementation of a policy (see chapter 7). We compare forecasted values with observed values to determine whether an intervention has the intended impact (see e.g. McCleary and Riggs 1982; Mohr 1992).

We can forecast future values by utilizing the information of past observations of a time series itself (see chapter 4) or by referring to the estimated relationship of the dependent time series variable with other time series variables (see chapter 6). In the latter case, we can forecast with greater accuracy, as the model's coefficient of determination is larger. However, we cannot know in advance what will be the exact values of predictor variables, even in the near future. Without the information of predictor variables in the future, forecasting with the estimated relationship of the dependent time series variable to other time series variables is only making a guess. In this case, forecasting with the estimated multiple time series analysis model could be worse than forecasting with the information of the past behavior of the dependent variable itself.

In addition, our multiple time series analysis model is generally not exhaustive. Future values forecasted by referring to the behavior of a time series itself may be more accurate than those forecasted by referring to the estimated relationship of the dependent time series variable with a few select independent variables. The behavior of the dependent time series variable will reflect influences from all factors that should be included in a multiple time series model.

However, our explanation of the forecast will be limited when we forecast future values by utilizing the information of past observations of a time series itself: we can provide explanations about our forecasts only in terms of the behavior of the time series but not in terms of factors that cause changes in the time series. Presidential-election outcomes, for example, can be influenced by various factors, such as the electorate's positions on salient issues, the state of the economy, characteristics of the candidates, presidential popularity, and evaluation of presidential job performance. With multiple time series analysis, we can provide explanations of our forecasts in terms of the relationships between these factors and presidential-election outcomes.

When we forecast future values by referring to the behavior of a time series itself, systematic patterns in residuals are important components of the time-dependent behavior of the time series. Without estimating such systematic patterns, our model is not complete and we cannot accurately forecast future values. For example, when we forecast future values with a time series with no discernible trend, as in figure 3, the trend line that is estimated with OLS regression analysis does not convey meaningful information. Even in this case, we may still identify underlying systematic patterns in residuals. Different systematic patterns will require different interpretations of the behavior of a time series and lead to different forecasted values. Figure 3, for example, can be interpreted in two different ways (Norpoth 1995). First, if there is no


FIGURE 3. Republican percentage of major-party vote for president, 1860–1992.
SOURCE: Norpoth (1995), Figure 1, p. 202. Reproduced with permission of Cambridge University Press.

systematic pattern, that is if the time series consists of independent and identically distributed noise,[2] we can say that the chance of winning the election is equally divided between the two political parties. To put it differently, there is perfect competition between the two parties. Second, if some type of dependency exists among the observations, that is if there is an autoregressive pattern in the series, we can say that one party tends to have advantage over the other for a certain period of time, and then the relationship is reversed for another period of time. By identifying an autoregressive model, we can estimate when the reversal of electoral advantage tends to occur.

In univariate time series analysis, we can estimate seasonality and systematic patterns of residuals and thereby improve the accuracy of our model and forecast. However, short-term regularities in residuals may be artifacts of factors that are not considered in univariate time series analysis. To take this possibility into consideration, we may employ multiple time series analysis that includes these factors (see chapter 6). The public approval of the president, for example, tends to increase for a certain period of time and to decrease for another period of time. However, this may be not because the public approval of the president is serially correlated but because economic conditions, such as unemployment and inflation, are serially correlated. If those factors show random fluctuations, presidential popularity may also randomly change.

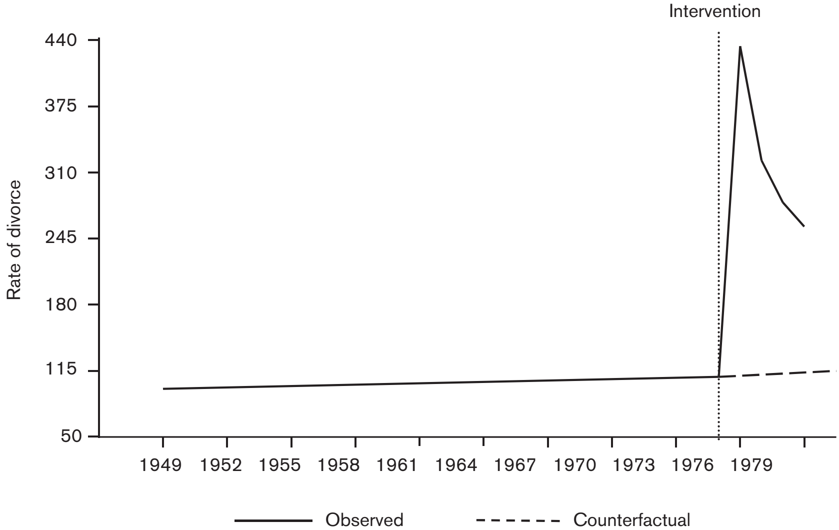


FIGURE 4. The 1975 Family Law Act and divorce rates in Australia.
SOURCE: Adapted from McCleary and Riggs (1982), Figure 1, p. 11. Reproduced with permission of John Wiley & Sons Inc.

Time series analysis is frequently employed in the impact analysis of a policy. Time series analysis used for this purpose is called *interrupted time series analysis*. The basic idea of interrupted time series analysis is that, if a program has an intended impact on the behavior of a target, the pre- and post-intervention segments of the temporal observations of the behavior should show significantly different levels and/or trends (Cook and Campbell 1979; McDowall et al. 1980; Mohr 1992). To put it differently, if a program has any significant impact, such an impact will be identified, as in figure 4, by the difference between the observations after the intervention point and the *counterfactual*, the projection of the correctly-modeled-before series into the post-intervention period (see chapter 7).

We can also test the impact of one or more factors on the dependent variable by comparing the observed values of the dependent variable with the values forecasted by the factors. For example, Mackuen et al. (1989) show that the observed macropartisanship—the percentage of Democrats divided by the percentage of Democrats plus Republicans—matches well the one predicted by presidential approval and consumer sentiment, meaning that the macropartisanship has been shaped by these short-term forces. In order to show that party identification is more stable than Mackuen et al. argue, Green,

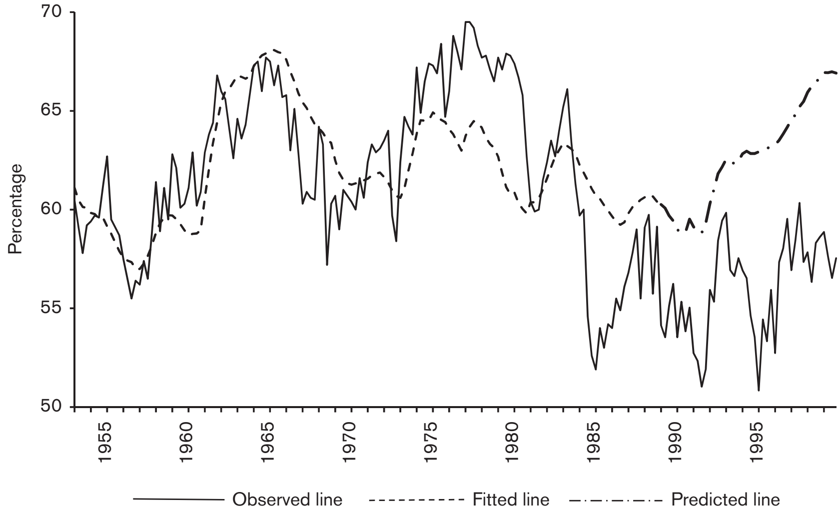


FIGURE 5. Observed and forecasted macropartisanship.

SOURCE: Adapted from Green et al. (2001), Figure 19-3, p. 362. Reproduced with permission of CQ Press.

Palmquist, and Schickler (2001) compared the observed macropartisanship with the forecasted one (figure 5). The forecast was based on a model that includes presidential approval, along with the lagged dependent variable and the systematic pattern of residuals, thereby controlling out the time-dependent behavior of the dependent variable. Green et al. show that when the estimated model is applied to new data points, macropartisanship is not well forecasted with presidential approval and thereby conclude that macropartisanship is stable to a greater extent than suggested by Mackuen et al.

Sometimes a variable is observed for a number of cross-sections over a time span. Combining cross-sectional data and time series data is useful when the number of available time points is too small and/or the sample of cross-sections is small in size (Sayrs 1989). The characteristics of data may also require us to employ a pooled cross-sectional analysis. Markus (1988), for example, includes in his model one individual-level variable (personal financial change) and one aggregate-level variable (the annual rate of change in real disposable personal income per capita). The latter can be cross-sectionally observed only in a comparative study of multiple nations. Markus, however, analyzes the presidential vote in only one nation, the United States. Thus, in order to estimate the impact of the aggregate-level variable on the presidential vote, he decides to cover a long period, from 1956 to 1984.

Time series analysis will continue to be a major interest in various social science disciplines. The purpose of this book is to provide easy-to-understand, practical guidelines for time series analysis to social scientists who want to understand studies, such as those cited above, that utilized time series analysis techniques or who want to conduct their own research and practice.