

# A Diagnostic for Assessing the Influence of Cases on the Prediction of Missing Data

Joseph E. Cavanaugh and Jacob J. Oleson

Department of Statistics, University of Missouri – Columbia

## Abstract

An important aspect of statistical modeling involves the identification of cases that have a significant impact on certain inferential results. In modeling problems where data is missing, the predicted values for the missing observations are frequently of interest (cf. Little and Rubin, 1987). To assist in the identification of cases that substantially influence these predicted values, we introduce a case-deletion diagnostic which is often conveniently evaluated in the setting of the EM algorithm (Dempster, Laird, and Rubin, 1977). Our diagnostic is defined as the Kullback-Leibler information (Kullback, 1968, p. 5) between two versions of the conditional density of the missing data given the observed data: one based on the parameter estimates arising from the full data set, the other based on the parameter estimates arising from the case-deleted data set. We outline the computation of the diagnostic for two Gaussian frameworks: for bivariate data applications in which some of the data pairs are incomplete, and for time series forecasting applications in which the missing observations are future realizations of the series. Our analyses involve bivariate data from the 1998 American Major League Baseball season and a time series consisting of cardiovascular mortality readings from the Los Angeles area.

**Keywords:** Case-deletion diagnostic, EM algorithm, forecasting, influence diagnostic, predictive influence function, time series.

**Corresponding Author:** Joseph E. Cavanaugh, Department of Statistics, 222 Mathematical Sciences Building, University of Missouri, Columbia, MO 65211. E-mail: cavanaugh@stat.missouri.edu.

## 1. Introduction

An important aspect of statistical modeling involves the identification of cases that have a significant influence on certain inferential results. Influence diagnostics have been extensively studied in linear regression, where measures such as Cook's distance, DFBETAS, DFFITS, and COVRATIO have gained widespread popularity. (See Cook and Weisberg, 1982; Belsley, Kuh, and Welsch, 1980.) The purpose of these diagnostics is to gauge the impact of a case on a specific inferential objective. Such measures compare inferential quantities (e.g., regression parameter estimates, fitted values, estimated generalized variances) based on fitting a model to the full data set with those based on fitting a model to the data set with a specific case removed. For this reason, such measures are often called case-deletion diagnostics.

In modeling problems that involve missing data, two goals are commonplace: to estimate the model parameters in the presence of the missing data, and to impute reasonable values for the missing observations. For settings where the latter goal is of fundamental interest, we propose a diagnostic that assesses the influence of a case on the prediction of the missing entries. Our diagnostic is defined as the Kullback-Leibler information (Kullback, 1968, p. 5) between two versions of the conditional density of the missing data given the observed data: one based on the parameter estimates arising from the full data set, the other based on the parameter estimates arising from the case-deleted data set. The diagnostic is often conveniently evaluated in the setting of the EM algorithm. The motivation for the diagnostic arises from the work of Johnson and Geisser (1983), Johnson (1985), and Cavanaugh and Johnson (1999).

Our paper is organized as follows. In Section 2, the diagnostic is introduced and two general computational formulae are presented for its evaluation. In Section 3, the implementation of the first formula is described for applications that involve incomplete bivariate normal data. The performance of the diagnostic is illustrated in an analysis featuring data from the 1998 American Major League Baseball season. In Section 4, the implementation of the second computational formula is outlined for applications that involve time series forecasting. An analysis is presented which considers cardiovascular mortality readings from the Los Angeles area. Section 5 concludes.

## 2. The Predictive Influence Function

Consider a statistical modeling application where a data set  $Y$  consists of missing values. Let  $Y_{obs}$  denote the observed part of  $Y$  and let  $Y_{mis}$  denote the missing part.

Suppose the complete data  $Y$  consists of a collection of  $n$  cases  $(y_1, \dots, y_n)$ . In some applications (such as the one considered in Section 3), each case will be a vector which may be comprised of both missing and observed components. The cases may be either correlated (see Section 4) or uncorrelated (see Section 3).

Let  $\theta$  denote the parameter vector for the model to be fit to the data. Assume  $\theta$  is to be estimated using the method of maximum likelihood, perhaps via the EM algorithm. Let  $\hat{\theta}$  represent the estimates based on all of the cases, and let  $\hat{\theta}^i$  represent the estimates based on all of the cases except case  $i$ .

The conditional density that dictates the prediction of the missing data is  $f(Y_{mis} | Y_{obs}, \hat{\theta})$ . One might therefore judge the influence of the  $i^{th}$  case on the prediction of  $Y_{mis}$  by measuring the extent to which the inclusion or the exclusion of this case affects  $f(Y_{mis} | Y_{obs}, \hat{\theta})$ : i.e., by measuring the disparity between  $f(Y_{mis} | Y_{obs}, \hat{\theta})$  and  $f(Y_{mis} | Y_{obs}, \hat{\theta}^i)$ . To gauge this disparity, we choose the Kullback-Leibler information.

We define the *predictive influence function* for assessing the impact of case  $i$  on the prediction of  $Y_{mis}$  as

$$\text{PIF}(i) = \int \left[ \log \left\{ \frac{f(Y_{mis} | Y_{obs}, \hat{\theta})}{f(Y_{mis} | Y_{obs}, \hat{\theta}^i)} \right\} \right] f(Y_{mis} | Y_{obs}, \hat{\theta}) dY_{mis}. \quad (2.1)$$

It is well-known that (2.1) is nonnegative. Moreover, the magnitude of (2.1) will reflect the divergence of  $f(Y_{mis} | Y_{obs}, \hat{\theta}^i)$  from  $f(Y_{mis} | Y_{obs}, \hat{\theta})$ .

The evaluation of  $\text{PIF}(i)$  is accomplished through one of the the following two formulas. To present these formulas, let  $L(\theta | Y)$  denote the complete-data likelihood and let  $L(\theta | Y_{obs})$  denote the incomplete-data likelihood. Define the functions  $Q(\theta | \theta_*)$  and  $H(\theta | \theta_*)$  as follows:

$$Q(\theta | \theta_*) = \int \{\log L(\theta | Y)\} f(Y_{mis} | Y_{obs}, \theta_*) dY_{mis}, \quad (2.2)$$

$$H(\theta | \theta_*) = \int \{\log f(Y_{mis} | Y_{obs}, \theta)\} f(Y_{mis} | Y_{obs}, \theta_*) dY_{mis}. \quad (2.3)$$

The function  $Q(\theta | \theta_*)$  is a familiar tool used in the implementation of the EM algorithm. The function  $H(\theta | \theta_*)$  can be interpreted as a measure of discrepancy between the densities  $f(Y_{mis} | Y_{obs}, \theta)$  and  $f(Y_{mis} | Y_{obs}, \theta_*)$ .

We have the following results, the first of which is justified in the Appendix, and the second of which follows directly from (2.1) and (2.3).

*Proposition.*

$$\text{PIF}(i) = \{Q(\hat{\theta} | \hat{\theta}) - Q(\hat{\theta}^i | \hat{\theta})\} + \{\log L(\hat{\theta}^i | Y_{obs}) - \log L(\hat{\theta} | Y_{obs})\}. \quad (2.4)$$

$$\text{PIF}(i) = H(\hat{\theta} | \hat{\theta}) - H(\hat{\theta}^i | \hat{\theta}). \quad (2.5)$$

The computation of  $\text{PIF}(i)$  via (2.4) is often convenient in missing-data applications where the EM algorithm is employed. With many EM implementations, the evaluation of both  $Q(\theta | \theta_*)$  and  $L(\theta | Y_{obs})$  is straightforward, even though the maximization of  $L(\theta | Y_{obs})$  with respect to  $\theta$  may be inconvenient or problematic. In these instances, the four terms comprising (2.4) are accessible once the EM algorithm is used to obtain the MLEs  $\hat{\theta}$  and  $\hat{\theta}^i$ .

The computation of  $\text{PIF}(i)$  via (2.5) is possible in missing-data applications where  $H(\theta | \theta_*)$  may be easily reduced and evaluated in terms of  $Y$ ,  $\theta$ , and  $\theta_*$ . For such settings, the EM algorithm may still be useful as a means for obtaining quantities that appear in the reduced formula: in particular, the MLEs  $\hat{\theta}$  and  $\hat{\theta}^i$ , and the predictors for the elements of  $Y_{mis}$ .

We now outline the evaluation of  $\text{PIF}(i)$  in two Gaussian frameworks. In Section 3 the computational formula involves the four terms in (2.4), whereas in Section 4 the formula is based on the two terms in (2.5).

### 3. Incomplete Bivariate Normal Applications

#### 3a. Evaluation of the Diagnostic

Suppose that the complete data  $Y$  consists of a collection of  $n$  data pairs  $(y_1, \dots, y_n)$ , where each case  $y_i$  is a  $2 \times 1$  vector of the form  $y_i = (y_{1i}, y_{2i})'$ ,  $i = 1, \dots, n$ . Assume that in some of the data pairs, an element is missing. The data pairs will be modeled as independent, identically distributed realizations of a bivariate normal distribution. The parameters of this

distribution are given by  $\theta = (\mu_1, \mu_2, \sigma_{11}, \sigma_{22}, \sigma_{12})'$ , where

$$\mu_1 = E(y_{1i}), \quad \mu_2 = E(y_{2i}), \quad \sigma_{11} = \text{Var}(y_{1i}), \quad \sigma_{22} = \text{Var}(y_{2i}), \quad \sigma_{12} = \text{Cov}(y_{1i}, y_{2i}).$$

The evaluation of  $\text{PIF}(i)$  via (2.4) requires the sufficient statistics for the parameters in  $\theta$ :

$$S_1 = \sum_{i=1}^n y_{1i}, \quad S_2 = \sum_{i=1}^n y_{2i}, \quad S_{11} = \sum_{i=1}^n y_{1i}^2, \quad S_{22} = \sum_{i=1}^n y_{2i}^2, \quad S_{12} = \sum_{i=1}^n y_{1i}y_{2i}.$$

In the preceding formulae, if an element  $y_{ji}$  ( $j = 1, 2$ ) of a data pair is missing, the element and its square must be replaced by their optimal predictors,  $E\{y_{ji} \mid Y_{obs}\}$  and  $E\{y_{ji}^2 \mid Y_{obs}\}$ . Specifically, if  $y_{1i}$  is missing in case  $i$ , it is replaced in  $S_1$  and  $S_{12}$  by

$$\tilde{y}_{1i} \equiv \mu_1 + \frac{\sigma_{12}}{\sigma_{22}}(y_{2i} - \mu_2),$$

and its square is replaced in  $S_{11}$  by

$$\tilde{y}_{1i}^2 + \left( \sigma_{11} - \frac{\sigma_{12}^2}{\sigma_{22}} \right).$$

If  $y_{2i}$  is missing in case  $i$ , it is replaced in  $S_2$  and  $S_{12}$  by

$$\tilde{y}_{2i} \equiv \mu_2 + \frac{\sigma_{12}}{\sigma_{11}}(y_{1i} - \mu_1),$$

and its square is replaced in  $S_{22}$  by

$$\tilde{y}_{2i}^2 + \left( \sigma_{22} - \frac{\sigma_{12}^2}{\sigma_{11}} \right).$$

Of course, the predictors for the missing values and their squares depend on the parameters of the model; thus, so do the complete-data sufficient statistics. To emphasize this dependence, we will denote the sufficient statistics by  $S_1(\theta), S_2(\theta), S_{11}(\theta), S_{22}(\theta), S_{12}(\theta)$ .

The first term in (2.4) is given by

$$\begin{aligned} Q(\hat{\theta} \mid \hat{\theta}) &= -n \log 2\pi - \frac{1}{2}n \log(\hat{\sigma}_{11}\hat{\sigma}_{22} - \hat{\sigma}_{12}^2) \\ &\quad - \frac{1}{2}(\hat{\sigma}_{11}\hat{\sigma}_{22} - \hat{\sigma}_{12}^2)^{-1} \left[ \hat{\sigma}_{22}S_{11}(\hat{\theta}) + \hat{\sigma}_{11}S_{22}(\hat{\theta}) - 2\hat{\sigma}_{12}S_{12}(\hat{\theta}) \right. \\ &\quad \left. - 2 \left\{ S_1(\hat{\theta}) (\hat{\mu}_1\hat{\sigma}_{22} - \hat{\mu}_2\hat{\sigma}_{12}) + S_2(\hat{\theta}) (\hat{\mu}_2\hat{\sigma}_{11} - \hat{\mu}_1\hat{\sigma}_{12}) \right\} \right. \\ &\quad \left. + n \left( \hat{\mu}_1^2\hat{\sigma}_{22} + \hat{\mu}_2^2\hat{\sigma}_{11} - 2\hat{\mu}_1\hat{\mu}_2\hat{\sigma}_{12} \right) \right], \end{aligned} \tag{3.1}$$

where  $\hat{\mu}_1, \hat{\mu}_2, \hat{\sigma}_{11}, \hat{\sigma}_{22}, \hat{\sigma}_{12}$  are the maximum likelihood estimates based on the data set consisting of all the cases: i.e.,  $(\hat{\mu}_1, \hat{\mu}_2, \hat{\sigma}_{11}, \hat{\sigma}_{22}, \hat{\sigma}_{12})' = \hat{\theta}$ . (See McLachlan and Krishnan, 1997, p. 47.) It may be shown that (3.1) reduces to

$$Q(\hat{\theta} | \hat{\theta}) = -n \log 2\pi - \frac{1}{2}n \log(\hat{\sigma}_{11}\hat{\sigma}_{22} - \hat{\sigma}_{12}^2) - n.$$

The second term in (2.4),  $Q(\hat{\theta}^i | \hat{\theta})$ , is computed in the same manner as (3.1), except that  $\hat{\mu}_1, \hat{\mu}_2, \hat{\sigma}_{11}, \hat{\sigma}_{22}, \hat{\sigma}_{12}$  are replaced with the maximum likelihood estimates based on the data set consisting of all the cases except case  $i$ : i.e.,  $(\hat{\mu}_1^i, \hat{\mu}_2^i, \hat{\sigma}_{11}^i, \hat{\sigma}_{22}^i, \hat{\sigma}_{12}^i)' = \hat{\theta}^i$ . Note that both  $Q(\hat{\theta} | \hat{\theta})$  and  $Q(\hat{\theta}^i | \hat{\theta})$  utilize the sufficient statistics associated with the fit of the model based on the full data set.

The third term in (2.4) is the incomplete-data empirical log-likelihood  $\log L(\hat{\theta} | Y_{obs})$ . This term is computed by adding the contributions for each of the  $n$  cases separately. These  $n$  cases can be split into three groups of data pairs. The first group consists of pairs where  $y_{1i}$  is missing and  $y_{2i}$  is observed; the second group consists of pairs where  $y_{1i}$  is observed and  $y_{2i}$  is missing; the third group consists of pairs where both  $y_{1i}$  and  $y_{2i}$  are observed.

For cases in the first and the second groups, where only one element is observed, the log-likelihood contribution is defined in terms of the univariate normal distribution based on only the observed portion of the data pair. For cases in the first group, the contribution is of the form

$$-\frac{1}{2} \log 2\pi - \frac{1}{2} \log \hat{\sigma}_{22} - \frac{1}{2} \left\{ \frac{(y_{2i} - \hat{\mu}_2)^2}{\hat{\sigma}_{22}} \right\}, \quad (3.2)$$

and for cases in the second group, the contribution is of the form

$$-\frac{1}{2} \log 2\pi - \frac{1}{2} \log \hat{\sigma}_{11} - \frac{1}{2} \left\{ \frac{(y_{1i} - \hat{\mu}_1)^2}{\hat{\sigma}_{11}} \right\}. \quad (3.3)$$

For cases in the third group, where both elements are observed, the log-likelihood contribution is defined in terms of the bivariate normal distribution and is of the form

$$\begin{aligned} & -\log 2\pi - \frac{1}{2} \log(\hat{\sigma}_{11}\hat{\sigma}_{22} - \hat{\sigma}_{12}^2) \\ & - \frac{1}{2}(\hat{\sigma}_{11}\hat{\sigma}_{22} - \hat{\sigma}_{12}^2)^{-1} \left[ \hat{\sigma}_{22}y_{1i}^2 + \hat{\sigma}_{11}y_{2i}^2 - 2\hat{\sigma}_{12}y_{1i}y_{2i} \right. \\ & \quad \left. - 2 \{ y_{1i}(\hat{\mu}_1\hat{\sigma}_{22} - \hat{\mu}_2\hat{\sigma}_{12}) + y_{2i}(\hat{\mu}_2\hat{\sigma}_{11} - \hat{\mu}_1\hat{\sigma}_{12}) \} \right. \\ & \quad \left. + (\hat{\mu}_1^2\hat{\sigma}_{22} + \hat{\mu}_2^2\hat{\sigma}_{11} - 2\hat{\mu}_1\hat{\mu}_2\hat{\sigma}_{12}) \right]. \end{aligned} \quad (3.4)$$

The sum of the log-likelihood contributions (3.2), (3.3), and (3.4) over all  $n$  cases yields  $\log L(\hat{\theta} | Y_{obs})$ . We emphasize that  $\hat{\mu}_1, \hat{\mu}_2, \hat{\sigma}_{11}, \hat{\sigma}_{22}, \hat{\sigma}_{12}$  are the maximum likelihood estimates based on the full data set: i.e.,  $(\hat{\mu}_1, \hat{\mu}_2, \hat{\sigma}_{11}, \hat{\sigma}_{22}, \hat{\sigma}_{12})' = \hat{\theta}$ . The same approach is used to find the fourth term in (2.4),  $\log L(\hat{\theta}^i | Y_{obs})$ , except that in (3.2), (3.3), and (3.4),  $\hat{\mu}_1, \hat{\mu}_2, \hat{\sigma}_{11}, \hat{\sigma}_{22}, \hat{\sigma}_{12}$  are replaced with the maximum likelihood estimates based on the data set with case  $i$  removed: i.e.,  $(\hat{\mu}_1^i, \hat{\mu}_2^i, \hat{\sigma}_{11}^i, \hat{\sigma}_{22}^i, \hat{\sigma}_{12}^i)' = \hat{\theta}^i$ .

We now illustrate the application of the diagnostic to a bivariate data set from the 1998 American Major League Baseball season.

### *3b. An Application*

We begin by defining some terms and concepts in baseball which are pertinent to our application.

The premise of baseball is to score more runs than the opposition. A run is scored when a player successfully touches all four bases before his team acquires three outs. The bases are 90 feet apart and form a diamond.

A player begins at home plate (the 4th base) holding a bat. The opposing team's pitcher throws the baseball to the batter who attempts to hit the ball with the bat. If the batter swings at a pitch and misses, a *strike* is called. If he elects not to swing at a pitch, a strike is called if the baseball's trajectory passes through a well-defined strike zone, and a *ball* is called otherwise. If the batter accumulates three strikes, he is called out; if he accumulates four balls, he is given a *walk* and advances to first base. If the batter hits the baseball into fair territory, he runs towards first base. Any teammates on base may also advance at this time. The batter is called out if his baseball is caught before it hits the ground. A player advancing towards a base, i.e., a runner, is called out if the baseball reaches the base before he does.

Whenever a batter hits a baseball that results in a run being scored, the batter is credited with a *run batted in* (RBI). If the batter hits the baseball out of the ballpark into fair territory, he is credited with a *home run*. In this instance, the batter along with each of his teammates currently on base automatically advances to home plate. The batter is then credited with

between one and four RBIs, depending on the number of teammates on base. Intuitively, a fairly strong positive correlation should exist between the number of home runs and the number of RBIs earned by a player over a certain time period, since both indices reflect the overall ability of a player at bat.

From the 1998 American Major League Baseball season, we examine a collection of 105 data pairs consisting of home runs and RBIs for players from the National League. In this season, the home run record held since 1961 by Roger Maris was broken by both Mark McGwire of the St. Louis Cardinals and Sammy Sosa of the Chicago Cubs.

Only players who had at least 350 at bats were included in the data set; all such players from the National League are represented. (An *at bat* is defined as the number of times a player bats, excluding when the player receives a walk, is hit by a pitched ball, or hits a sacrifice.) By eliminating players who do not satisfy the aforementioned criterion, we greatly attenuate the degree of right skewness that is inherent in both the home run and RBI data sets, thereby making the assumption of bivariate normality more reasonable.

Letting  $y_{1i}$  denote the number of home runs and  $y_{2i}$  the number of RBIs for player  $i$ , we assume that the data pairs  $(y_{1i}, y_{2i})$  can be at least approximately described by the bivariate normal model. To make the data set incomplete, we randomly discard 3 home run entries (for cases 45, 58, and 94) and 3 RBI entries (for cases 3, 8, and 88). By artificially introducing missing values in a data set which is fully observed, we are then able to check the accuracy of the predicted values for the missing elements against the actual entries.

The EM algorithm is used to compute maximum likelihood estimates of  $\mu_1, \mu_2, \sigma_{11}, \sigma_{22}, \sigma_{12}$  for the full data set. Predicted values are found for the missing observations using the resulting MLEs. The EM algorithm is then used to compute MLEs for the data set with the  $i^{th}$  case removed,  $i = 1, \dots, 105$ . Predicted values are subsequently found using the case-deleted MLEs. PIF values are finally calculated using (2.4).

The plot of the PIF( $i$ ) against the index  $i$  is provided in Figure 1. Based on the magnitude of PIF(32), case 32 appears to be highly influential in the prediction of the missing values. The vast majority of the PIF values are less than 1% the size of PIF(32), including all the values associated with the incomplete cases.

The following table features the parameter estimates for the full data set along with the estimates for the data set with case 32 removed. Note that the deletion of case 32 reduces both variances as well as the covariance. The correlation is left relatively unchanged.

Data Set	$\mu_1$	$\mu_2$	$\sigma_{11}$	$\sigma_{22}$	$\sigma_{12}$
All Cases	18.48	73.27	162.44	785.77	316.16
All Cases Except 32	17.97	72.57	137.44	745.15	282.69

The scatterplot of the 105 data pairs is presented in Figure 2. Dots are used to designate cases where both entries are observed; circles are used to indicate incomplete cases. Case 32 is marked with a diamond. Note that this case appears to lie below the linear trend established by the remainder of the data.

The following table features the predicted values for the missing entries based on both the fitted model for the full data set and the fitted model for the data set with case 32 removed. The predicted values are rounded to the nearest integer. The actual entries are also featured for the purpose of comparison. Clearly, the deletion of case 32 results in a fitted model that predicts the missing observations with greater overall accuracy. (Note that the mean square error of prediction is 350 based on the full data set and 232 based on the data set without case 32.)

Missing Home Run Entry

Case	Actual Value	Predicted Value:	
		All Cases	All Cases Except 32
45	2	0	1
58	31	33	32
94	23	24	23

Missing RBI Entry

Case	Actual Value	Predicted Value:	
		All Cases	All Cases Except 32
3	39	45	44
8	45	49	48
88	144	127	130

Case 32 represents the RBIs and home runs for Mark McGwire. McGwire broke a 37-year-old home run record by hitting 9 more home runs than had ever been hit before in a single season. McGwire led the major leagues in walks as well, setting a National League record for the most walks in a single season. When McGwire was at bat with runners in scoring position, the opposing team would often choose to deliberately walk him and take its chances with the next batter. Thus, many of McGwire’s home runs resulted in solo scoring. This caused McGwire’s RBIs to be lower than expected given the number of home runs he hit. Hence, in a plot of RBIs versus home runs (Figure 2), McGwire’s case falls below the trend established by the remainder of the data. Additionally, his case has high leverage, thereby influencing the parameter estimates and consequently the predicted values for the missing entries.

Sammy Sosa also broke the home run record during the 1998 season: Sosa hit 66 home runs as opposed to McGwire’s 70. One may therefore suspect that Sosa’s case is influential. Sosa’s case corresponds to the index 19. Note in Figure 1 that  $PIF(19)$  is not unusually large; thus, the deletion of case 19 should not substantially impact the prediction of the missing values. In fact, when rounded to the nearest integer, the predicted values for cases 3, 8, 45, 58, and 94 are exactly the same for the full data set as they are for the data set with case 19 deleted, and the predicted value for case 88 differs only by a factor of one (127 versus 128).

Case 19 is marked by a triangle in Figure 2; this case also has high leverage, yet appears better aligned with the overall trend than case 32.

Sosa was not walked as often as McGwire when there were runners in scoring position: in fact, Sosa was walked 89 fewer times than McGwire over the course of the season. This allowed Sosa to accrue more RBIs than McGwire. As a result, Sosa’s RBIs are more consistent with the number of home runs he hit.

## 4. Gaussian Time Series Forecasting Applications

### *4a. Evaluation of the Diagnostic*

Consider a univariate Gaussian time series of length  $T$  represented by  $(y_1, \dots, y_T)$ . Suppose our objective is to model the time series parametrically, and to use the fitted model to

produce the  $h$ -step forecast  $y_{T+h}$ ,  $h \geq 1$ .

In the present context, we can regard  $Y_{mis}$  as the point to be forecast,  $y_{T+h}$ . The observed data is  $Y_{obs} = (y_1, \dots, y_T)$ .

We will use  $\theta$  to denote the model parameters. The forecast (best predictor) for  $y_{T+h}$  is

$$E\{y_{T+h} \mid Y_{obs}\} \equiv y_{T+h}^T(\theta),$$

and the conditional variance for the forecast is

$$E\{(y_{T+h} - y_{T+h}^T(\theta))^2 \mid Y_{obs}\} \equiv P_{T+h}^T(\theta).$$

In the Appendix, we derive the following computational formula for  $\text{PIF}_h(i)$ , which assesses the effect of the deletion of case  $i$  on the forecast for  $y_{T+h}$ :

$$\begin{aligned} \text{PIF}_h(i) = & \left\{ -\frac{1}{2} \log P_{T+h}^T(\hat{\theta}) - \frac{1}{2} \right\} \\ & - \left\{ -\frac{1}{2} \log P_{T+h}^T(\hat{\theta}^i) - \frac{1}{2} (P_{T+h}^T(\hat{\theta}) / P_{T+h}^T(\hat{\theta}^i)) \right. \\ & \left. - \frac{1}{2} (y_{T+h}^T(\hat{\theta}) - y_{T+h}^T(\hat{\theta}^i))^2 / P_{T+h}^T(\hat{\theta}^i) \right\}. \end{aligned} \quad (4.1)$$

Note that the two bracketed terms in the preceding respectively correspond to  $H(\hat{\theta} \mid \hat{\theta})$  and  $H(\hat{\theta}^i \mid \hat{\theta})$  in (2.5).

In many applications, our interest may lie in forecasting a sequence of future values, say  $y_{T+1}$  through  $y_{T+L}$ . To obtain a diagnostic that reflects the impact of case  $i$  on this entire set of forecasts, we may simply add together the PIFs for each of the individual forecasts: i.e.,

$$\text{PIF}(i) = \sum_{h=1}^L \text{PIF}_h(i). \quad (4.2)$$

An alternative approach consists of developing a PIF where the missing data is regarded as  $Y_{mis} = (y_{T+1}, \dots, y_{T+L})$ . This leads to a formula that is analogous to (4.1), yet one which requires the evaluation of the covariances between forecasts. The formula is therefore more computationally cumbersome to evaluate than (4.2).

We now illustrate the application of the diagnostic to a time series consisting of cardiovascular mortality readings.

#### 4b. An Application

Figure 3 features a plot of cardiovascular mortality readings recorded for the Los Angeles area during the 1970s. Each reading represents an average taken over a 6-day period, so that the 180 displayed observations cover a time span of roughly 3 years. Note that yearly cycles are evident in the series.

If the series is modeled using the autoregressive moving-average (ARMA) framework, a simple autoregressive model of order two appears to provide an adequate fit of the data:

$$y_t = \alpha + \phi_1 y_{t-1} + \phi_2 y_{t-2} + w_t; \quad w_t \sim N(0, \sigma^2); \quad t = 1, \dots, 180.$$

For the fitted AR(2) model, Ljung-Box (1978) tests do not detect autocorrelations in the residuals, and the residuals appear to be approximately normally distributed.

In the ARMA framework, case-deleted MLEs may be found by representing the model in state-space form, and by either applying the EM algorithm or numerically maximizing the innovations form of the likelihood. (See, respectively, Shumway and Stoffer, 1982; Jones, 1980.) Since the evaluation of the complete-data tool  $Q(\theta | \theta_*)$  is not required in (4.1), we use the latter approach. The  $h$ -step forecasts  $y_{T+h}^T(\theta)$  and their associated variances  $P_{T+h}^T(\theta)$  may be determined simply by extending the Kalman filter recursions for  $h$  steps beyond the end of the series. (See Brockwell and Davis, 1991, pp. 477-482.)

We consider the influence of each case on the forecasts for  $y_{181}$  through  $y_{186}$ . PIF values are evaluated using (4.2) (with  $T = 180; L = 6; i = 1, \dots, 180$ ).

The plot of the PIF( $i$ ) against the index  $i$  is provided in Figure 4. Based on the magnitude of PIF(77), case 77 appears to be highly influential in the forecasting of the next 6 realizations of the series. This case is marked with an asterisk in Figure 3. Note that this observation corresponds to an unusually high spike that appears during the low part of the cycle for the second year.

The following table features the forecasts for  $y_{181}$  through  $y_{186}$  based on the fitted AR(2) model for both the full data set and the data set with case 77 removed. For the purpose of comparison, the table also features the forecasts based on the fitted AR(2) model for the

data set with case 20 removed. Note from Figure 4 that PIF(20) is quite small; we would therefore suspect that the deletion of this case has little impact on the forecasts.

Time Point	Forecast: All Cases	Forecast: Case 20 Deleted	Forecast: Case 77 Deleted
181	83.8	83.8	83.6
182	84.9	84.9	84.6
183	85.8	85.8	85.4
184	86.6	86.6	86.2
185	87.3	87.3	86.9
186	88.0	88.0	87.5

Whereas the removal of case 20 has no effect on the forecasts, the deletion of case 77 reduces the magnitude of each predicted value. This decrease becomes more extreme as the forecasts move further into the future.

The following table features the AR(2) model parameter estimates for the full data set along with the estimates for the data set with case 77 removed. Note that the deletion of case 77 reduces the estimates for both the mean adjustment  $\alpha$  and the error variance  $\sigma^2$ . The former leads to the decrease in the forecasts; the latter leads to a decrease in the forecast variances. Both types of reductions affect the relevant forecast densities: i.e., the conditional densities of the  $y_{180+h}$  ( $h = 1, \dots, 6$ ) given  $Y_{obs}$ . This results in the large value of the PIF for case 77.

Data Set	$\alpha$	$\phi_1$	$\phi_2$	$\sigma^2$
All Cases	14.47	0.350	0.496	38.98
All Cases Except 77	13.20	0.342	0.517	36.36

Clearly, changes in predicted values result from changes in the fitted model. Thus, one may question whether PIF tends to flag the same cases as more conventional case-deletion diagnostics, specifically those designed to detect shifts in parameter estimates. To investigate this issue, we employ several such diagnostics in the present application.

First, we consider three diagnostics which are analogous to the DFβETAS used in linear regression. Let  $\mu = E(y_t)$  (i.e.,  $\alpha = \mu - \phi_1\mu - \phi_2\mu$ ), and let  $SE(\cdot)$  represent an estimated standard error for an MLE. We define

$$\text{DFMU} = \frac{\hat{\mu} - \hat{\mu}^i}{\text{SE}(\hat{\mu}^i)}, \quad \text{DFAR1} = \frac{\hat{\phi}_1 - \hat{\phi}_1^i}{\text{SE}(\hat{\phi}_1^i)}, \quad \text{DFAR2} = \frac{\hat{\phi}_2 - \hat{\phi}_2^i}{\text{SE}(\hat{\phi}_2^i)}.$$

Second, we consider two diagnostics introduced by Bruce and Martin (1989). Let  $\phi = (\phi_1, \phi_2)'$ , let  $\Sigma(\cdot)$  denote the large-sample variance/covariance matrix for a vector of MLEs, and let  $n$  denote the sample size. The first diagnostic, which assesses changes in the estimate of the error variance, is defined as

$$\text{DV} = (n/2) \left( \frac{(\hat{\sigma}^2)}{(\hat{\sigma}^2)^i} - 1 \right)^2.$$

The second, which reflects changes in the estimates of the autoregressive parameters, is defined as

$$\text{DC} = n(\hat{\phi} - \hat{\phi}^i)' \Sigma(\hat{\phi})^{-1} (\hat{\phi} - \hat{\phi}^i).$$

The values of  $\text{PIF}(i)$  featured in Figure 4 show that the magnitude of the diagnostic for case 77 is much larger than that for any other case. This is not true for any of the estimate-based influence measures.

With  $\text{DFMU}$ ,  $\text{DFAR1}$ ,  $\text{DFAR2}$ , and  $\text{DC}$ , no individual case produces an extreme value. Moreover, case 77 does not produce the largest value for any of these diagnostics. It is worth noting that all of the values of the  $\text{DFBETAS}$  measures are well below 0.5: thus, no single case deletion shifts a parameter estimate more than half a standard error.

With  $\text{DV}$ , the value for case 77 is the largest; this is not surprising since the deletion of case 77 has a greater impact on the estimate of the error variance than on the estimates of the other model parameters. However, cases 91 and 151 also have unusually large values of  $\text{DV}$ .

The preceding results imply that an analysis based on case-deletion diagnostics which assess changes in parameter estimates may fail to reflect the influence of case 77 on the forecasts. The results are not paradoxical, since the forecasts (and forecast variances) produced by a model may be more sensitive to certain perturbations in the model coefficients than to others. Hence, the deletion of a case may have a more pronounced impact on the parameter estimates for the fitted model than it has on the forecasts produced by the fitted model, or vice versa.

## 5. Conclusion

In statistical modeling, it is important to identify cases that have a substantial impact on key inferential results. Such cases may indicate recording errors or anomalies in the phenomenon that produced the data. Such cases may also serve as an indication that the underlying model is too simplistic; thus, the problems of model selection and influential case detection must be addressed jointly.

The applications in Section 3 and 4 illustrate that the predictive influence function is effective in flagging cases which impact the prediction of missing values. Such cases are often not apparent from a visual inspection of the data.

We note that the diagnostic could easily be developed for many modeling frameworks where the EM algorithm is used. This not only includes applications in which the missing data is missing in the conventional sense, but also applications in which the missing data represents unobservable quantities that are routinely predicted: e.g., random effects in mixed models, latent factors in factor analysis models, etc.

We also note that the diagnostic could be used with either single or multiple case deletion. The latter approach could be beneficial to identify possible masking effects (i.e., where the influence of one case is obscured by the presence of another case), or to detect influential “patches” in time series analysis (cf., Bruce and Martin, 1989).

In future work, we hope to derive a baseline to determine when a value of the diagnostic is “large.” Our initial investigations, which utilize results from Shimodaira (1994) and Cavanaugh and Shumway (1997), suggest that such a baseline would depend on the amount of missing information present in the application, and may require quantities obtained from the SEM algorithm (Meng and Rubin, 1991).

## Acknowledgements

The authors wish to thank Simon Davies, J. Wade Davis, Thomas Bengtsson, and Ryan Murphy for their comments and suggestions. The work of the first author was supported by the National Science Foundation, grant DMS-9704436.

## Appendix

### *Proof of (2.4) in Proposition of Section 2*

Note that

$$\log f(Y_{mis} | Y_{obs}, \theta) = \log L(\theta | Y) - \log L(\theta | Y_{obs}). \quad (\text{A.1})$$

Using (A.1) along with the definition (2.2), we have

$$\int \{\log f(Y_{mis} | Y_{obs}, \hat{\theta})\} f(Y_{mis} | Y_{obs}, \hat{\theta}) dY_{mis} = Q(\hat{\theta} | \hat{\theta}) - \log L(\hat{\theta} | Y_{obs}), \quad (\text{A.2})$$

$$\int \{\log f(Y_{mis} | Y_{obs}, \hat{\theta}^i)\} f(Y_{mis} | Y_{obs}, \hat{\theta}) dY_{mis} = Q(\hat{\theta}^i | \hat{\theta}) - \log L(\hat{\theta}^i | Y_{obs}). \quad (\text{A.3})$$

Expression (2.4) then follows from utilizing (A.2) and (A.3) in conjunction with the definition of PIF( $i$ ) provided by (2.1).  $\square$

### *Derivation of (4.1)*

The complete-data likelihood  $L(\theta | Y)$  may be factored into a product of the incomplete-data likelihood  $L(\theta | Y_{obs})$  and the conditional density of  $y_{T+h}$  given  $Y_{obs}$ . This yields

$$\begin{aligned} \log L(\theta | Y) &= \log L(\theta | Y_{obs}) \\ &\quad - \frac{1}{2} \log 2\pi - \frac{1}{2} \log P_{T+h}^T(\theta) - \frac{1}{2} (y_{T+h} - y_{T+h}^T(\theta))^2 / P_{T+h}^T(\theta). \end{aligned} \quad (\text{A.4})$$

Let  $E_*(\cdot | Y_{obs})$  denote the expectation operator with respect to the conditional density  $f(y_{T+h} | Y_{obs}, \theta_*)$ . With reference to (A.4), we have

$$\begin{aligned} Q(\theta | \theta_*) &= E_* \{ \log L(\theta | Y) | Y_{obs} \} \\ &= \log L(\theta | Y_{obs}) - \frac{1}{2} \log 2\pi - \frac{1}{2} \log P_{T+h}^T(\theta) \\ &\quad - \frac{1}{2} E_* \left[ \left\{ (y_{T+h} - y_{T+h}^T(\theta_*)) + (y_{T+h}^T(\theta_*) - y_{T+h}^T(\theta)) \right\}^2 | Y_{obs} \right] / P_{T+h}^T(\theta) \\ &= \log L(\theta | Y_{obs}) - \frac{1}{2} \log 2\pi - \frac{1}{2} \log P_{T+h}^T(\theta) \\ &\quad - \frac{1}{2} E_* \left\{ (y_{T+h} - y_{T+h}^T(\theta_*))^2 | Y_{obs} \right\} / P_{T+h}^T(\theta) - \frac{1}{2} (y_{T+h}^T(\theta_*) - y_{T+h}^T(\theta))^2 / P_{T+h}^T(\theta) \\ &\quad - \frac{1}{2} (y_{T+h}^T(\theta_*) - y_{T+h}^T(\theta)) E_* \left\{ (y_{T+h} - y_{T+h}^T(\theta_*)) | Y_{obs} \right\} / P_{T+h}^T(\theta) \\ &= \log L(\theta | Y_{obs}) - \frac{1}{2} \log 2\pi - \frac{1}{2} \log P_{T+h}^T(\theta) \\ &\quad - \frac{1}{2} (P_{T+h}^T(\theta_*) / P_{T+h}^T(\theta)) - \frac{1}{2} (y_{T+h}^T(\theta_*) - y_{T+h}^T(\theta))^2 / P_{T+h}^T(\theta). \end{aligned} \quad (\text{A.5})$$

The computational formula for PIF $_h(i)$ , (4.1), then follows by using (A.5) to evaluate  $Q(\hat{\theta} | \hat{\theta})$  and  $Q(\hat{\theta}^i | \hat{\theta})$ , and by substituting these expressions into (2.4).  $\square$

## References

- Belsley, D. A., Kuh, E. and Welsch, R. E. (1980). *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. New York: John Wiley and Sons.
- Brockwell, P. J. and Davis, R. A. (1991). *Time Series: Theory and Methods (Second Edition)*. New York: Springer-Verlag.
- Bruce, A. G. and Martin, R. D. (1989). Leave-k-out diagnostics for time series. *Journal of the Royal Statistical Society, Series B* **51**, 363–401.
- Cavanaugh, J. E. and Johnson, W. O. (1999). Assessing the predictive influence of cases in a state-space process. *Biometrika* **86**, 183–190.
- Cavanaugh, J. E. and Shumway, R. H. (1997). An Akaike information criterion for model selection in the presence of incomplete data. *Journal of Statistical Planning and Inference* **67**, 45–65.
- Cook, R. D. and Weisberg, S. (1982). *Residuals and Influence in Regression*. London: Chapman and Hall.
- Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B* **39**, 1–38.
- Johnson, W. (1985). Influence measures for logistic regression: Another point of view. *Biometrika* **72**, 59–65.
- Johnson, W. and Geisser, S. (1983). A predictive view of the detection and characterization of influential observations in regression analysis. *Journal of the American Statistical Association* **78**, 137–144.
- Jones, R. H. (1980). Maximum likelihood fitting of ARMA models to time series with missing observations. *Technometrics* **22**, 389–395.
- Kullback, S. (1968). *Information Theory and Statistics*. New York: Dover.
- Little, R. J. A. and Rubin, D. B. (1987). *Statistical Analysis with Missing Data*. New York: John Wiley and Sons.
- Ljung, G. M. and Box, G. E. P. (1978). On a measure of lack of fit in time series models. *Biometrika* **65**, 297–304.

- McLachlan, G. J. and Krishnan, T. (1997). *The EM Algorithm and Extensions*. New York: John Wiley and Sons.
- Meng, X. L. and Rubin, D. B. (1991). Using EM to obtain asymptotic variance-covariance matrices: The SEM algorithm. *Journal of the American Statistical Association* **86**, 899–909.
- Shimodaira, H. (1994). A new criterion for selecting models from partially observed data. In: P. Cheeseman and R. W. Oldford, Eds., *Selecting Models from Data: Artificial Intelligence and Statistics IV, Lecture Notes in Statistics* **89**, 21–29. New York: Springer-Verlag.
- Shumway, R. H. and Stoffer, D. S. (1982). An approach to time series smoothing and forecasting using the EM algorithm. *Journal of Time Series Analysis* **3**, 253–264.

Figure 1. PIF values for baseball data.

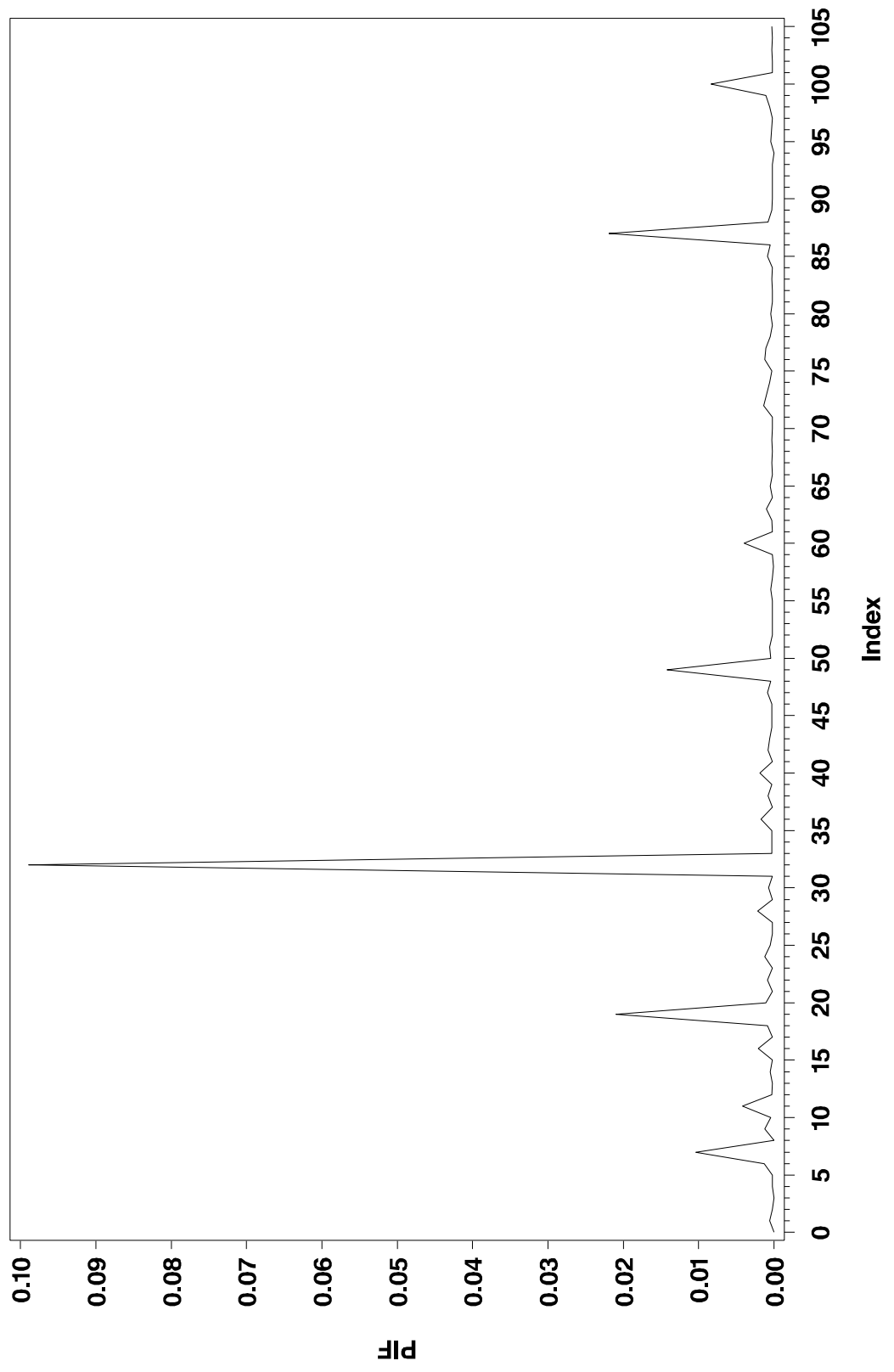


Figure 2. RBIs versus home runs for 105 National League players (1998 season).

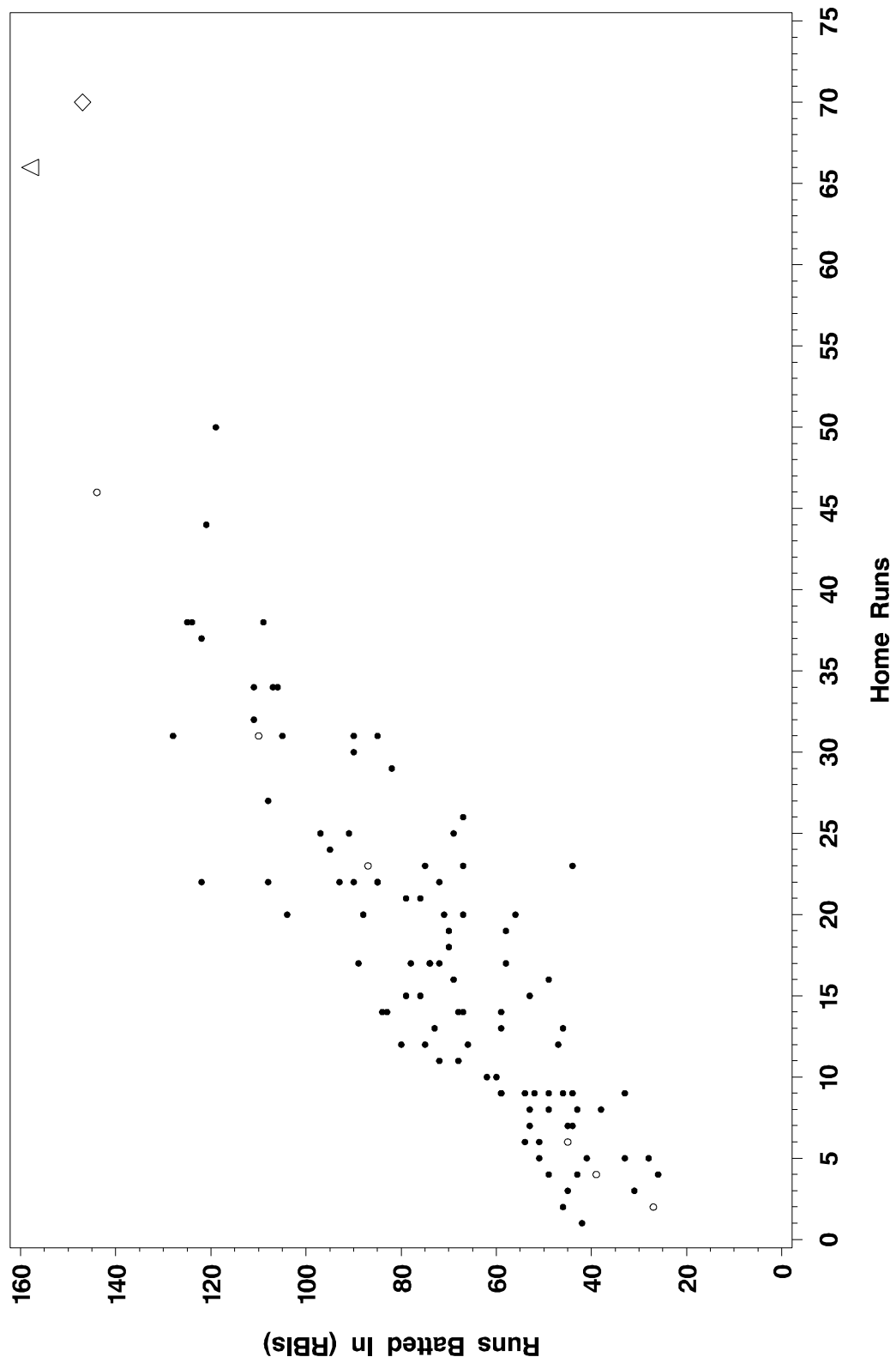


Figure 3. Los Angeles cardiovascular mortality data.

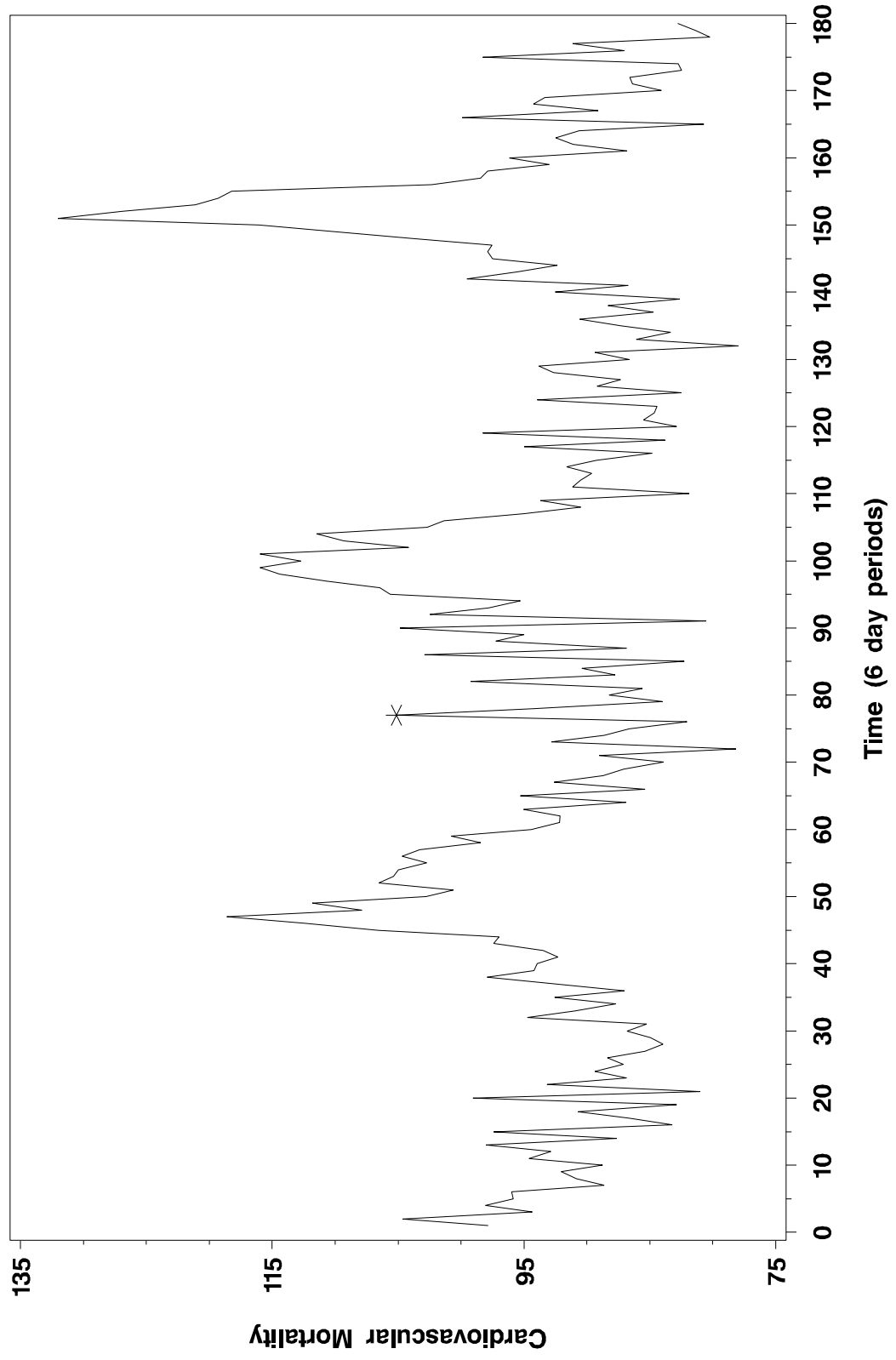


Figure 4. PIF values for Los Angeles cardiovascular mortality data.

