U.S. DEPARTMENT OF COMMERCE
NATIONAL OCEANIC AND ATMOSPHERIC ADMINISTRATION
NATIONAL WEATHER SERVICE
OFFICE OF SCIENCE AND TECHNOLOGY
METEOROLOGICAL DEVELOPMENT LABORATORY

MDL OFFICE NOTE 02-1

A METHODOLOGY FOR EVALUATING AND ESTIMATING PERFORMANCE METRICS

Harry R. Glahn

February 2002

A METHODOLOGY FOR EVALUATING AND ESTIMATING PERFORMANCE METRICS

Harry R. Glahn

1. INTRODUCTION

One of the ways the National Weather Service (NWS) evaluates its service is by computing metrics for a selected sample of forecasts. For instance the Probability of Detection (POD), False Alarm Ratio (FAR), and Critical Success Index (CSI) are computed for a number of event forecasts such as tornadoes and flash flood.[1] Other metrics, or verification scores, are mean absolute error (MAE) and root mean square error (RMSE) for continuous variables such as temperature and wind speed. Mean square error (MSE) is computed for the probability of precipitation (PoP) and is usually called the Brier Score. The lead time in forecasting certain events is also computed, for example, for tornadoes, flash flood, and hurricane landfall.

Generally, these metrics are summarized and presented in terms of monthly, seasonal, or yearly values. Because most metrics are quite dependent on time of year (and time of day if relevant), tracking over time is done in a way that the seasonal variability does not dominate and distract from the main issue at hand. For instance, plotting MAEs for summer seasons will present a coherent pattern that will indicate how the accuracy level is changing, if at all; scores for different seasons plotted on one chart are harder to comprehend. In a like manner, yearly scores have the seasonal influence masked, but provide a good overall picture of the accuracy or skill of the forecasts to which the scores pertain.

Even when the diurnal and seasonal aspects are not dealt with, either by averaging them out, or dealing with one specific time of day or year, there is considerable variability in scores due to the variability of weather patterns, accuracy of the verifying observations, and the variability in sampling that is done of the total number of forecasts made. So, when an additional score is added at the end of a year and it is above or below the previous year's score, the question arises as to whether this difference indicates an increase or decrease in skill, or whether it is just due to "natural" variability.

Many metrics computed on NWS forecasts have been showing improvement, most of then in fact. But improvement usually comes in small doses. Major technological improvements, like the introduction of Doppler radar and highly improved computing resources, have occasioned improvements in some forecasts that

[1] Definitions of metrics are contained in the <u>National Verification Plan</u> (1982). "Scores" and "metrics" are used interchangeably in this paper.

are immediately apparent, but these situations are the exception rather than the norm. So how do we judge when an improvement has been made?

Metrics can be plotted for a number of years, and the trend, if any, established from those scores. This usually does not address whether or not forecasts for a particular year were "better" or "worse," but rather whether or not there was general improvement over a number of years. On the other hand, a "new" score may be so different from the past scores as to indicate real change in performance. Some method is needed to assist the analyst in judging whether a change has occurred; that is, how much different does the score have to be to indicate a change not due to natural variability?

In addition to the need to assess past performance, the NWS sets goals for metrics for a few years in advance. Such metrics can only reasonably be established by analysis of past metrics. So then, the question is how to analyze past data to project the metrics into the future.

Usually, the number of past scores that can be used for prediction is rather limited, and this limits the complexity of analysis that can be legitimately done. For instance, the analysis of a half dozen scores is basically limited to a one or two parameter model. A one parameter model can be to project the mean (the one parameter) of the past scores into the future--forecast no change in the scores within the next few years. Or to account for possible trend, a two parameter linear regression model is appropriate. With the same data, confidence bands can be put on the regression "line" at some desired level of significance such that only the corresponding fraction of scores would fall outside those bounds merely by chance, provided the assumptions on which the model is based are true. Major departures can be judged suspect. In the same manner, the regression line can be extended into the near future, with the understanding that the line really applies to the data analyzed and not to future data. However, in the absence of better tools and the necessity to make such projections, the regression analysis provides an appropriate framework for making such estimates. As stated by Neter and Wasserman (1974) (hereafter called NW), pp. 29-30, **"Regression analysis serves three major purposes: (1) description, (2) control, and (3) prediction....The several purposes of regression analysis frequently overlap in practice."**

This paper presents a methodology that can be used in judging whether changes have occurred and for making estimates for future scores. This will be done in the framework of yearly scores, but can be applied to other situations, for instance the probability of detection of low ceiling heights at a particular hour in a particular month, given sufficient data on which to base the analysis. Examples are shown to illustrate the methodology.

2. LEAST SQUARES LINEAR REGRESSION

A. The Model

The two parameter linear regression model is very simple. It relates a predictand (also called the dependent variable) Y to a single predictor (also called the independent variable) X. It is written:

$$Y = a + bX$$

The parameters a and b can be estimated in various ways and with various assumptions, least squares being the most used. Y_i is treated as a random variable with constant variance at each value of X_i, regardless of the value of X_i (NW, p. 31) and X_i is a known constant (NW, p. 30).[2] This equation applied to the data points being analyzed yields an estimate of Y, called $\hat{Y}$, such that $\Sigma(Y_i - \hat{Y}_i)^2$ over all n data points is a minimum; no other line can say that.

The task is to find estimates of a and b. This is straightforward if the variance of all data points is the same. That is, there is no reason to believe the variation in the measurement (or computation) of Y_1 is different from the variation in the measurement of Y_2, etc. In contrast, the measurement error of Y could depend on X, and the model would not strictly apply.

The solution of the so-called "normal" equations provides the estimates of a and b (NW, p. 37):

$$b = \frac{\Sigma X_i Y_i - \Sigma X_i \Sigma Y_i / n}{\Sigma X_i^2 - (\Sigma X_i)^2 / n}$$

$$a = \bar{Y} - b\bar{X}$$

where

$$\bar{Y} = \Sigma Y_i/n,$$

$$\bar{X} = \Sigma X_i/n,$$

and the summations are over all n points.

[2] The model still applies when the X_i are random variables (NW, p. 76).

B. Example

Lead times for tornadoes (the time between when the warning was issued and the report of a tornado in the warned area) are available from post Doppler years 1995 through 2001[3]. These have been summarized into yearly mean lead times for the entire United States and are shown in the appendix. The linear regression model has been applied to these seven scores, and the regression line

$$Y = -131.0988 + 0.0708X$$

is plotted in Fig. 1 and extended to 2006. The reduction of variance (correlation coefficient r squared) is low, being only 0.053. One immediately wonders whether this is "significant." The question more formally stated is, can the null hypothesis

$$h_o: r^2 = 0$$

be rejected considering the alternative hypotheses

$$h_1: r^2 \neq 0?$$

The answer is resoundingly "No." The value to test with the t-test is only 0.53, and it would require a value of 2.01 for the null hypothesis to be rejected at the 90 percent level.[4] Does this mean the regression line cannot be used. Again the answer is "No." It just means we are not very sure of the location of the line, and the error bars (see below) will indicate that. This line, as unsure as we are of it, is all we have unless we want to use persistence (the most recent score) or the one parameter model where the mean of the scores is projected forward.

Also plotted in Fig. 1 are the confidence bands at α = 90, 95, and 99 percent levels. These confidence levels were computed according to (NW, p. 68):

$$\hat{Y}_h \pm t(1-\alpha/2;n-2)s(\hat{Y}_h),$$

where

$$Y_h = a + bX_h,$$

[3] Lead times for years prior to 1995 are available, but cannot reasonably be used without a more complicated analysis. For instance, piecewise regression (NW, p. 196), where the line changes at some point, could be used, but would not help in projecting the scores into the future.

[4] The significance test for the reduction of variance being different from zero is equivalent to the test for the slope of the line, b, being different from zero.

X_h is some specific value of X, t is the t distribution with n-2 degrees of freedom at the 1-α (two-tailed) level of significance, and $s(\hat{Y}_h)$ is the estimated standard deviation of $\hat{Y}_h$ given by the square root of the estimated variance

$$s^2(\hat{Y}_h) = MSE \left| \frac{1}{n} + \frac{(X_h - \bar{X})^2}{\Sigma(X_i - \bar{X})^2} \right| . \qquad \text{(NW, p. 67)}$$

$$MSE = \frac{\Sigma(Y_i - \hat{Y}_i)^2}{n - 2} \qquad \text{(NW, p. 45)}$$

where the MSE is the error (or residual or unexplained) sum of squares divided by the degrees of freedom, n-2.

These confidence bands apply to the <u>mean</u> response. This may be a bit confusing. If repeated samples could be taken at a given X, call it X_h, then the mean of those samples, the "expected" value of Y, $E(Y_h)$, would be expected to lie within the confidence bounds at that point the quoted percent of the time. However, we cannot take repeated samples of the yearly mean. To elaborate further, the model assumes each yearly mean lead time is drawn from a normally distributed sample of mean lead times, and the mean of those means is what the error bars apply to. Note that only one of the seven scores used in the analysis falls outside the 95 percent bands, not an unreasonable expectation.

Note that there is a confidence interval for each value of X_i, that it is a minimum at $\bar{X}$, and its width varies with $X_h - \bar{X}$. These individual values when plotted about the regression line and connected form hyperbolas [Draper and Smith, 1966, (hereafter called DS) p. 23].

From the above discussion, it is apparent the <u>confidence</u> bounds in Fig. 1 are not very useful for <u>prediction</u> because we cannot make multiple measurements of yearly mean scores. Fortunately, there is another way of computing bounds that <u>is</u> useful.

If a new score becomes available, (score, here, is the mean of several measurements, but not a mean of several means), then the estimator of the variance of the new score is

$$s^2(\hat{Y}_{h(new)}) = MSE \left| 1 + \frac{1}{n} + \frac{(X_h - \bar{X})^2}{\Sigma(X_i - \bar{X})^2} \right|$$

and the <u>prediction</u> bounds are computed in the manner given above

$$\hat{Y}_{h(new)} \pm t(1-\alpha/2;n-2)s(\hat{Y}_{h(new)}).$$

This wider interval accounts for the fact the confidence limits pertain to the mean response, while the prediction limits pertain to a single new independent score (NW, pp. 69-73 give a particularly good discussion; DS, pp. 23-24; Dixon and Massey 1983, p. 217). Fig. 2 indicates the prediction bands for the same regression line for an individual score.

A practical application of Fig. 2 is that if metric goals are to be set for this variable, mean tornado lead time, the regression line provides a good estimate, provided there are no major changes expected for the years for which the estimates are to be made that would materially affect the scores. That is, there is a small improvement indicated by the past scores, and this might be expected to continue. This small improvement may have been due to a number of small factors, and small factors, even different ones, can be expected to improve the scores in the future.

The prediction bounds indicate that if a goal were to be set above the upper 95 percent bound, then given the conditions stated above, one would expect the new score to fall above the bound at the new X only about 2.5 percent of the time. So if stretch goals are set and they are outside the upper confidence bound, then some dramatic change would likely be necessary to influence the future scores.

If a new score falls outside the 95 percent bound, one can be 95 percent sure the change is not due to random fluctuation in the scores. For instance, the "chance" component of the score should not cause it to fall below the lower confidence band more than 2.5 percent of the time.

3. WEIGHTED LEAST SQUARES LINEAR REGRESSION

A. The Model

The linear regression model discussed above is based on certain assumptions. Basic is the assumption that Y_i is treated as a random variable with constant variance at each value of X_i, regardless of the value of X_i (NW, p. 31). This essentially means that if there were a random variable, tornado lead time, with the same distribution as the observed tornado lead times, and repeated random samples were taken at each of the X_i, then the means of those samples would have the same variance. Although we are not really dealing with a random variable--we seldom are in meteorological analysis--the assumption seems reasonable. At least we know of nothing much better as an assumption. However, if we did know the sampling variability at each X_i, then we could use weighted least squares.

Estimates of a and b are then given by (Neter, et al., p. 419):

$$b = \frac{\Sigma w_i X_i Y_i - \Sigma w_i X_i \Sigma w_i Y_i / \Sigma w_i}{\Sigma w_i X_i^2 - (\Sigma w_i X_i)^2 / \Sigma w_i}$$

$$a = \Sigma w_i Y_i / \Sigma w_i - b \Sigma w_i X_i / \Sigma w_i$$

and the w_i should be the inverse of the variances of yearly scores. The trick is, what are the w_i?

In this case, two things present themselves as possibilities. First, the individual lead times are available (from which the yearly means were computed). The variance of the means, $\sigma_{\bar{X}}^2$, could be estimated from the variance of the individual lead times, and usually, according to the central limit theorem, this should provide a good estimate. However, the distribution of the individual lead times is <u>highly</u> non-normal, and the estimate of $\sigma_{\bar{X}}^2 = \sigma_X^2/n$ may not be a good estimate.

The other possibility is that the weight would vary with the number of cases on which each score was computed. This seems reasonable, and follows the suggestion by Montgomery and Peck, (1982, p. 99) who state, "In some problems, the weights may be easily determined. For example, if the observation Y_i is actually an average of n_i observations at X_i, and if all *original* observations have constant variance σ^2, then the variance of Y_i is σ^2/n_i, and we would choose the weights as $w_i = n_i$."

B. Example

Applying weighted least squares to the tornado lead time data used above and using the number of cases as weights gives the regression line:

$$Y = -197.9869 + 0.1043X$$

This line and the associated prediction bands are plotted in Fig. 3. The prediction bands are plotted in the same manner as above, where:

$$s^2(\hat{Y}_{h(new)}) = MSE \left| 1 + \frac{1}{n} + \frac{(X_h - \bar{X}_w)^2}{\Sigma(X_i - \bar{X}_w)^2} \right|$$

and $\bar{X}_w$ refers to the weighted X so that

$$\bar{X}_w = \Sigma w_i X / \Sigma w_i .$$[5]

[5] It is interesting to note that the mean of X is different than with unweighted least squares in addition to the mean of Y being different.

MSE is calculated the same way as before with the $\hat{Y}$ estimated from the weighted regression line:

$$MSE = \frac{\Sigma(Y_i - \hat{Y}_i)^2}{n - 2} \qquad \text{(NW, p. 45)}$$

Confidence bands for weighted least squares may not be as sure to contain the prescribed percentage of the data points if the weights are considerably different for different points. For instance, a point with a very small weight will not influence the placement of the line or the confidence bands very much and might lie outside the bands.

In a similar manner, with prediction, one would expect a new point to have a weight of the same order of magnitude as the weights used in the analysis for the prediction bands to apply. This just means the new score should be of the same quality as the scores on which the analysis was based.

This line and the line obtained with unweighted regression are compared in Fig. 4. Since all n_i are of the same order of magnitude, there is no practical difference in the two regression lines. Predictions made by one line would not vary much from predictions made by the other even out to 5 years in advance. The weighted regression line has a slightly larger slope primarily because the relatively low 2001 score is only for a partial year and the number of cases is smaller. Therefore, it had relatively less effect than it did with the unweighted solution.

4. LOGISTIC FUNCTION

A. The Model

There are situations where a metric is limited in scope. For instance, certain measurements are binary such as the forecasting of an event like flash flood. Given the occurrence, it was either forecast or not--a binary outcome. Means of the measurements (detection or not) give the relative frequency of detection and lie between 0 and 1. While each original measurement would be at either $y = 0$ or $y = 1$, the means would lie somewhere between, given repeated sampling at a given value of X (that is, the scores would be aggregated by year, say).

Linear regression, either weighted or unweighted can be used to estimate the relative frequency, or probability of detection POD, although it is easy to visualize that a linear line would be a poor fit, and projected forward in time could fall outside the 0-1 interval. A model appropriate for fitting such data is the logit:

$$Y = \frac{e^{(a + bX)}}{1 + e^{(a + bX)}}.$$

According to Cox (1989, p.19), "It will turn out that in many ways the most useful analogue for binary response data of the linear model for normally distributed data is provided by the linear logistic model."

This form can be used to solve for a and b with the original data, Y being binary. However, when multiple measurements are available at each X and none of the relative frequencies is 0 or 1, the transformation

$$P = \ln \left| \frac{\bar{Y}}{1 - \bar{Y}} \right|$$

"linearizes" the model, which then becomes

$$P = a + bX,$$

and a and b can be solved by regression (Montgomery and Peck, pp. 238-241).

It is obvious in this situation that the variance σ^2 is not constant for all p_i because the event is binary and σ^2 is different for the different values of relative frequency. For the linearized model when each n_i is large, the value of σ^2 is estimated to be:

$$\sigma^2 = 1/[n_i\bar{p}_i(1 - \bar{p}_i)]$$

and the weights are therefore (Montgomery and Peck, pp. 239-240)

$$w_i = n_i\bar{p}_i(1 - \bar{p}_i).$$

After solving for a and b, the estimates P_i can be transformed back to the original space by:

$$Y_i = \frac{e^{(a + bX_i)}}{1 + e^{(a + bX_i)}} = \frac{e^{(P_i)}}{1 + e^{(P_i)}}.$$

B. Example

No real data were readily available, so a dummy set, shown in the appendix, was used. Both the unweighted liner least squares regression and the weighted logit model were applied with all n_i equal, and the results are plotted in Figs. 5 and 6. This set of data represents what might be probabilities of detection, possibly of a winter storm 24 hours in advance. There is a good linear relationship shown over the past few years, and if linear regression were used to project into the future, impossible values of over 100 percent would be reached. On the other hand, the logit gives realistic results with error bars shown. The error bars were computed with the transformed data according to the weighted liner least squares regression model and transformed back to the original metric space.

5. DISCUSSION

Generally, least squares regression is a good model for determining whether a new value of a metric is within what would be expected by chance or whether it represents a possible real improvement or deterioration in performance as measured by the metric.

Many times the number of cases representing each value of the metric is large and does not vary much; in this case it matters little whether weighted or unweighted least squares regression is used.

When dealing with metrics which are relative frequencies, the logit model can be used when the values are near (but not extremely near) 0 or 1. Within values of about 0.2 and 0.8. and when the trend is small, the logit fit will be close to that for regression. It is only when the "prediction line" will approach 0 or 1 does the logit need be used.

A major consideration in any analysis of this sort is to not use data for which major abrupt changes have occurred. For instance, the introduction of a new numerical model might make such a change in the metric that a trend analysis would be inappropriate until a few scores were available for the new model. Also, observing practices could change, making the verifying observations of such a different character that data before and after the change should not be mixed.

There are instances where the metric is limited in scope other than the binary event discussed above. For instance, MAE for temperature forecasts may be quite low (good), and the recent trend may have been downward, but this trend cannot be expected to continue indefinitely. In such cases, projecting the last metric or the mean of the last few may be the most reasonable thing to do.

Another instance of limitation in scope is where the percent improvement is calculated. A favorite metric for the probability of precipitation is the improvement over climate. This will be capped at 100 percent. However, generally the scores are not close to 100 percent, and regression is adequate.

The software on which this paper is based is available from the author.

REFERENCES

Cox, D. R., and E. J. Snell, 1989: Analysis of Binary Data, Chapman and Hall, New York, 236 pp.

Dixon, W. J., and F. J. Massey, Jr., 1983: Introduction to Statistical Analysis, Fourth Edition. McGraw-Hill Book Company, New York, 678 pp.

Draper, N. R., and H. Smith, 1966: Applied Regression Analysis. John Wiley and Sons, Inc., New York, 407 pp.

Neter J., and W. Wasserman, 1974: Applied Linear Statistical Models. Richard D. Irwin, Inc., Homewood, Illinois, 842 pp.

Neter J., W. Wasserman, and M. H. Kutner, 1990: Applied Linear Statistical Models, Third Edition. Richard D. Irwin, Inc., Homewood, Illinois, 1181 pp.

NWS, 1982: National Verification Plan. National Weather Service, NOAA, U.S. Department of Commerce, 81 pp.

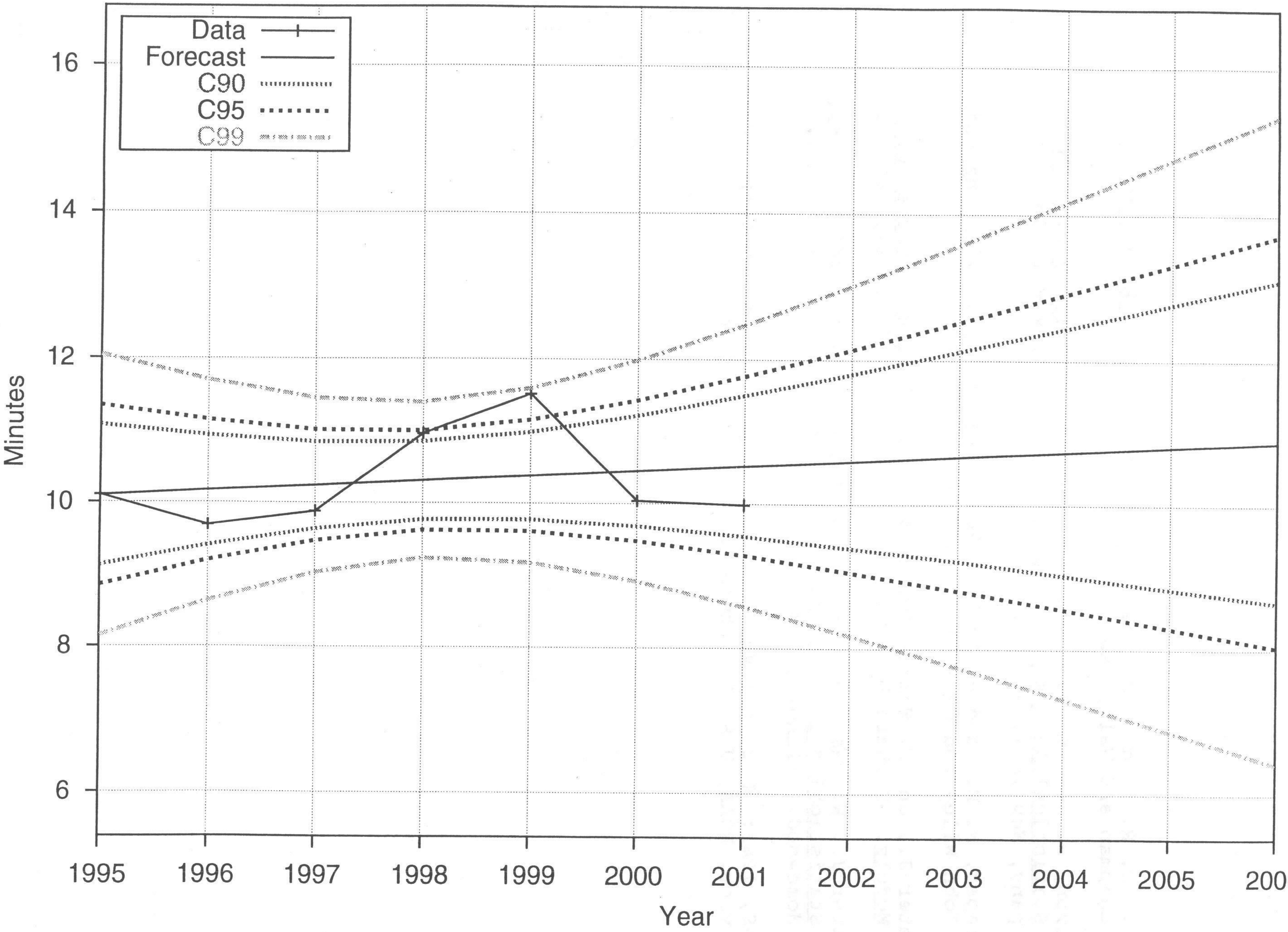

Figure 1. Regression line and confidence bounds at the 90, 95, and 99 percent levels for tornado lead time.

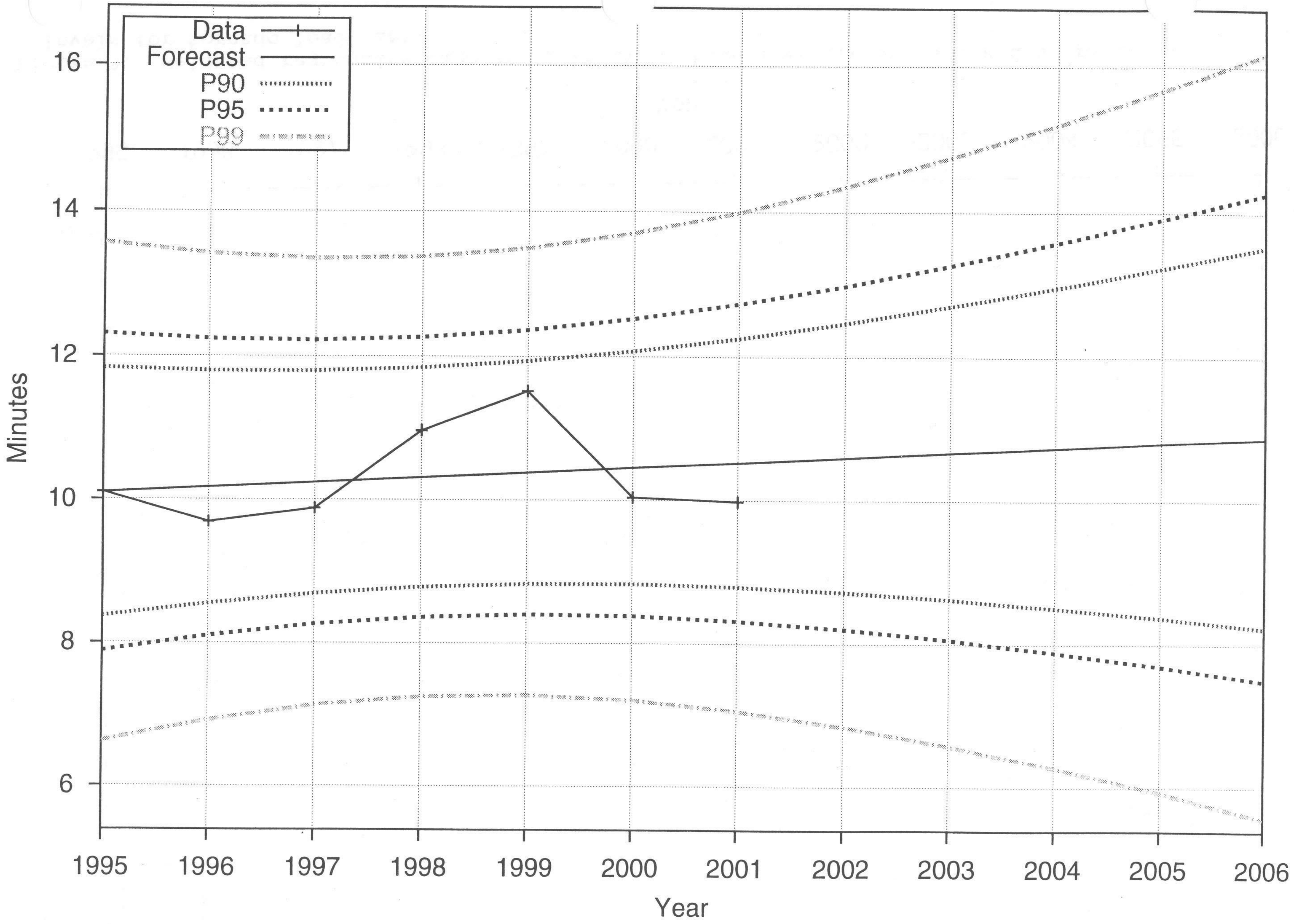

Figure 2. Regression line and prediction bounds at the 90, 95, and 99 percent levels for tornado lead time.

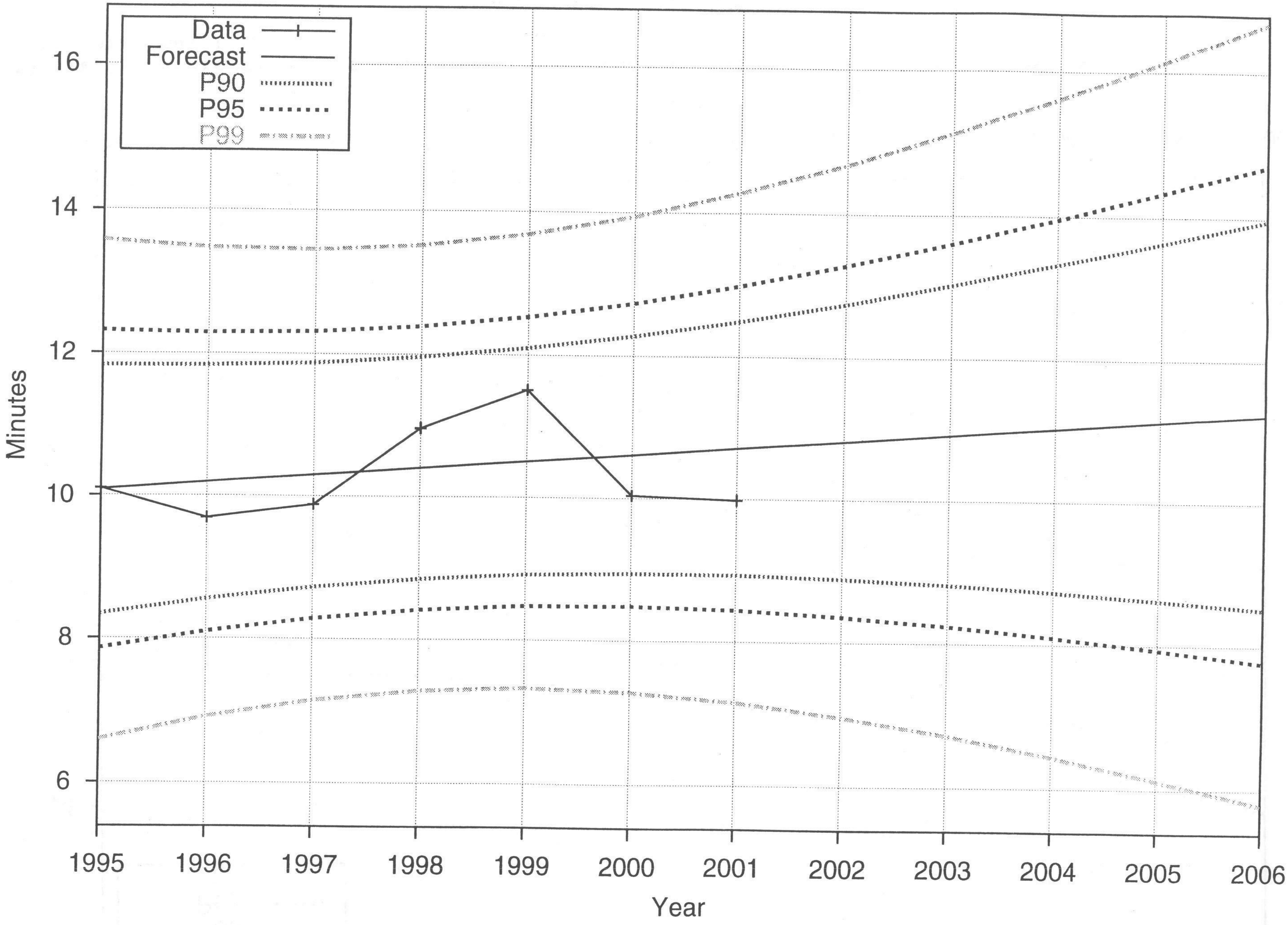

Figure 3. Weighted regression line and prediction bounds at the 90, 95, and 99 percent levels for tornado lead time.

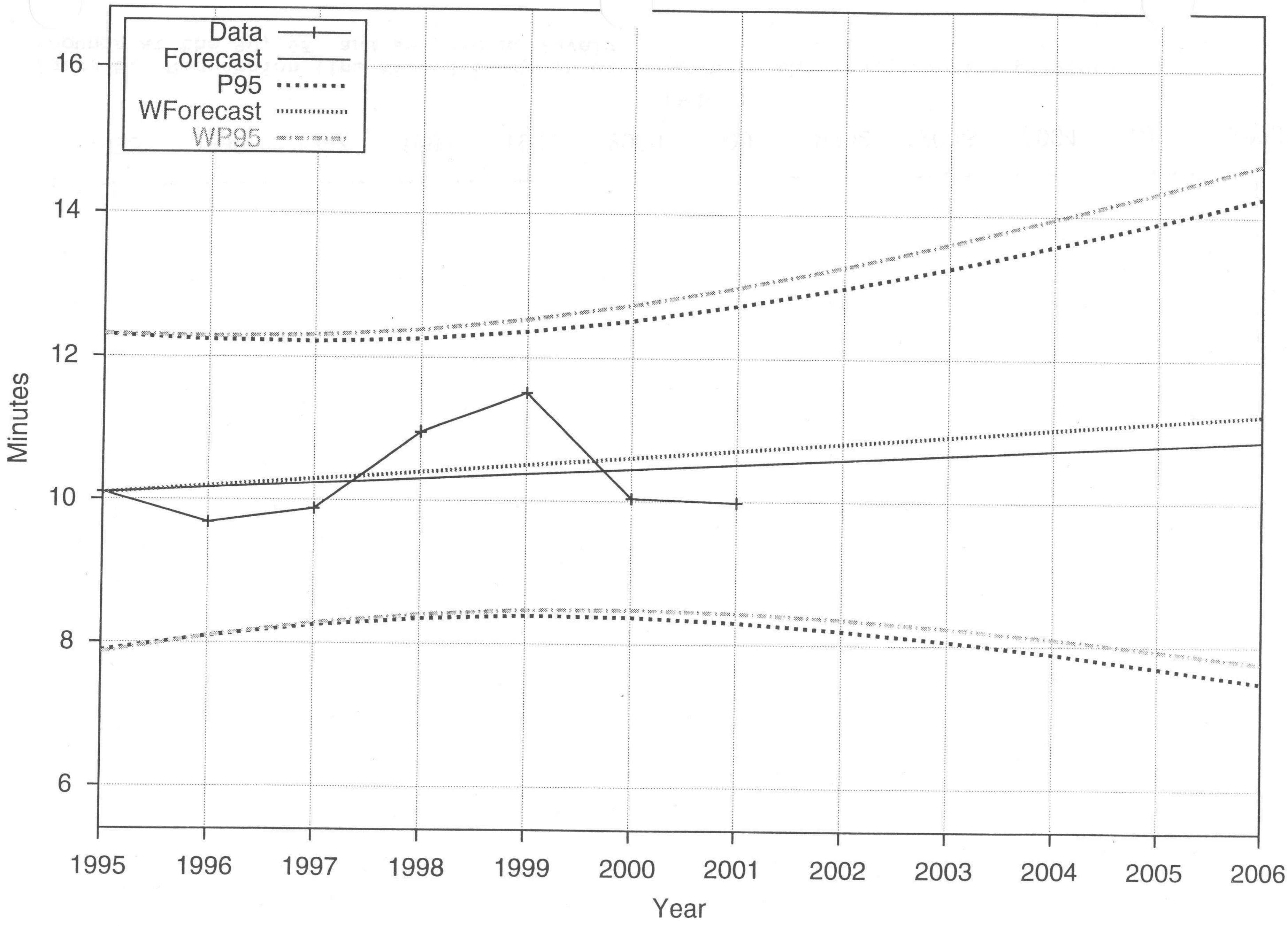

Figure 4. Comparison of weighted and unweighted regression lines and associated 95 percent prediction bounds.

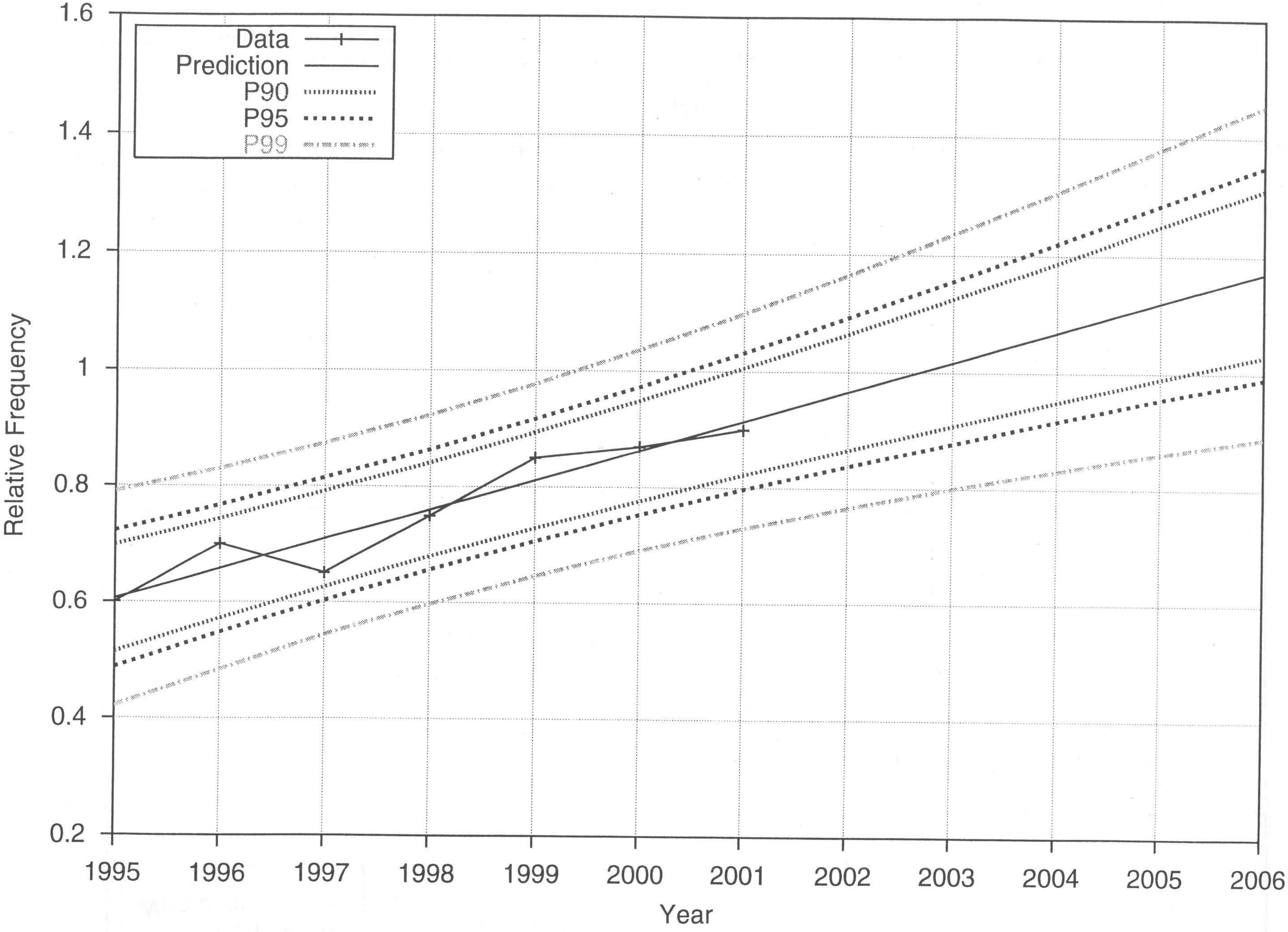

Figure 5. Regression line fitted to dummy data shown in the appendix with prediction bounds at the 90, 95, and 99 percent levels.

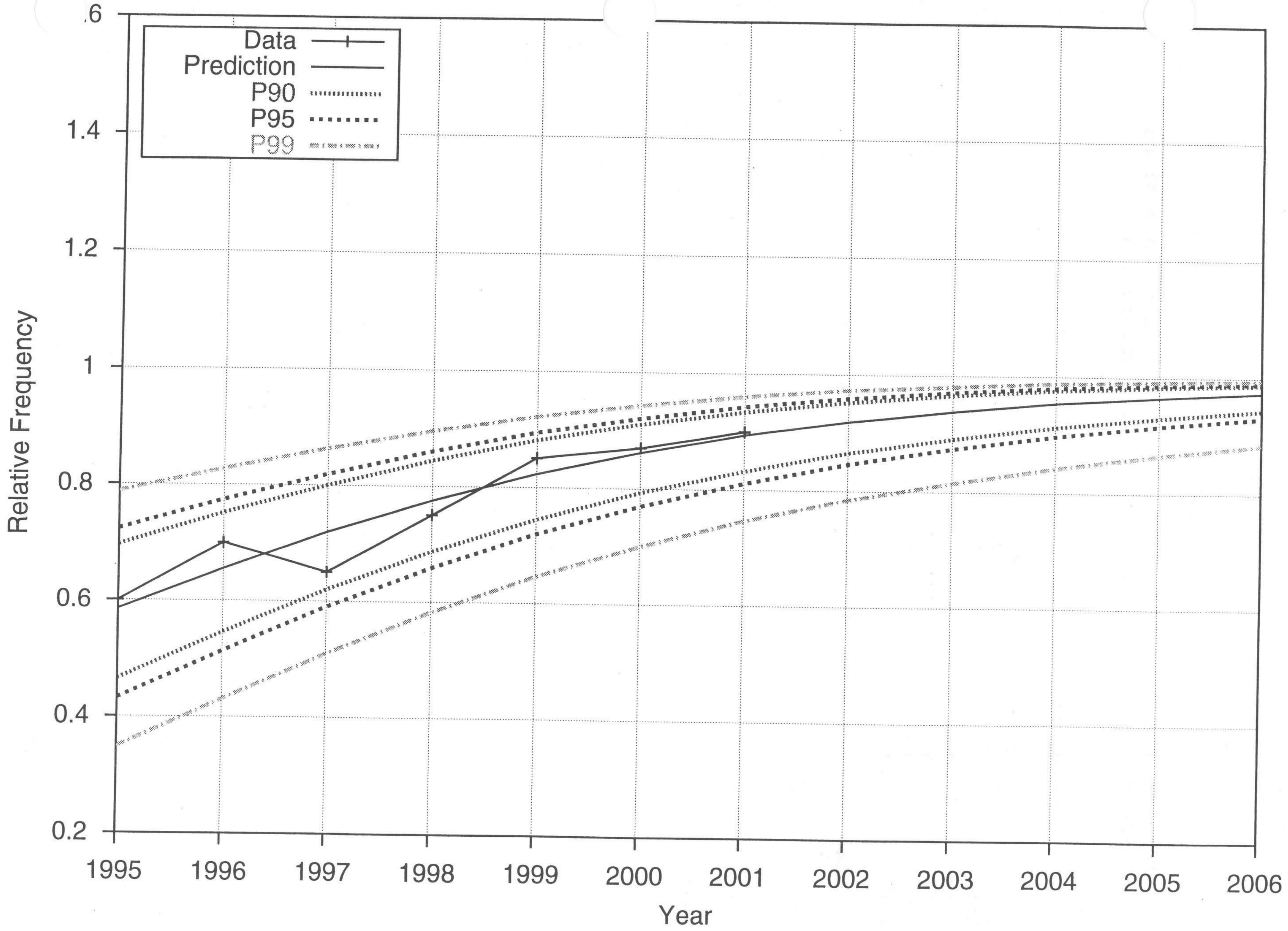

Figure 6. Similar to Fig. 5, except for the logit model.

APPENDIX

The tornado mean lead time data used in this paper are given below. The values for 2001 are not for a complete year.

Year	Lead Time (minutes)	No. of Cases
1995	10.1010	1297
1996	9.6888	1221
1997	9.8813	1163
1998	10.9606	1522
1999	11.5133	1505
2000	10.0381	1155
2001	9.9847	851

The Dummy data on which the logit model was based is:

Year	POD
1995	.60
1996	.70
1997	.65
1998	.75
1999	.85
2000	.87
2001	.90