

---

# Pitcher's Luck: Bringing the Computer in from the Bullpen

by

ROBERT M. JENNINGS

and

JOHN E. FINDLING

Indiana University Southeast

In recent years, computers have expanded greatly in the ability to process large amounts of information. In many cases, the processing has enabled users to draw meaningful inferences from the data. One method of analyzing the relationships indicated by the computer input is through the use of a multiple regression equation. This is a formidable term that requires some simplifying explanations. Essentially, a multiple regression equation permits the user to evaluate the proportionate impact of a number of independent variables acting upon a dependent variable.

For example, assume that over a ten-year career, a pitcher consistently won 65% of games he pitched on Tuesday and Friday, 55% of games pitched Wednesday and Thursday, and 45% of games pitched Monday and Saturday. (He never pitched on Sunday because of his religious beliefs.) An equation describing his wins in a year would be  $.65 (Tu \& Fr) + .55 (We \& Th) + .45 (Mo \& Sa)$ . It would be possible to then write a second equation describing his wins broken down by day or night games, a third equation by average field temperature above or below  $75^{\circ}$ , and so forth. In this example, then, a multiple regression equation could be formed by combining all of these into one computer program. The independent variables are those factors which might influence the pitcher's won-lost record (the dependent variable). What we learn from the multiple regression equation is which factors impact most strongly on the won-lost record and from that, we can predict a standard won-lost record and see if a pitcher's actual statistics exceed or fall short of the computer's prediction.

With no criteria other than pragmatism and some knowledge of baseball, the authors chose twenty pitchers with at least five years of substantial major league pitching. They then utilized the multiple regression computer program available through our university facility to predict the won-lost record for all the pitchers. The pragmatic nature of the factors that we assumed as independent variables is shown by the following list:

- 1) earned run average, coded  $x_1$
- 2) innings pitched, coded  $x_2$
- 3) team batting average, coded  $x_3$
- 4) team fielding average, coded  $x_4$
- 5) team runs per game, coded  $x_5$
- 6) opponents' runs per game, coded  $x_6$
- 7) team slugging average, coded  $x_7$
- 8) complete games, coded  $x_8$

As previously mentioned, these factors are the independent variables in mathematics vocabulary. The single predicted value (won-lost percentage) is the dependent variable. The rounded general equation, based on the sample of one hundred mentioned above, is:

$$-.0704x_1 - .00027x_2 - 1.797x_3 + 2.124x_4 + .0872x_5 \\ - .03076x_6 + .56877x_7 + .0061x_8 - 1.297$$

The standard error of the estimate was .084. Without going into statistical methods in any depth, we can state that the difference between the estimated value (in this case, the pitcher's won-lost percentage) and the actual value should be less than .165 about 95% of the time.

From the general equation, there are a number of interesting points of discussion (or conclusions to be drawn?). Are any pitchers consistently above their prediction (i.e., lucky)? Are any consistently below their prediction (i.e., unlucky)? How closely does this general equation describe performance of pitchers other than our sample of twenty? What other significant influencing factors might be included in our model?

In a subjective test of the accuracy of the general prediction equation, the authors chose nine cases that they felt were unusual enough to warrant examination on an individual basis. These cases were: Sandy Koufax (1963 and 1964); Whitey Ford (1961); Dick Radatz (1964); Jack Billingham (1973, 1974, and 1975); Roger Craig (1962 and 1963).

The program was then run to generate an estimate of the won-lost percentage of each of these cases, and that estimate was compared with the actual percentage. The difference could be either positive or negative. If the difference were positive (the actual percentage higher than the predicted percentage), then the pitcher could be termed "lucky" that year, and if the difference were negative, then the pitcher could be considered "unlucky."

Of the nine cases chosen and considering a 5% level of significance (i.e., a difference of plus or minus .165 as previously defined), only two were significant. These were Whitey Ford in 1961 and Roger Craig in 1963. Ford, whose score was .2058, could be viewed as "lucky," and Craig, whose score was -.236, could be described as "very unlucky."

To our knowledge, this is the first attempt to use the multiple regression equation to determine the relationship of a pitcher's won-lost percentage to some of its probable determinants. As such, it is a rough, rule-of-thumb, pioneer study.