Selection of Significant NBA Statistics for the Prediction of Wins
Semester Project
STAT 511
Dr. Naomi Altman

by: Eli Walters and Brian Staudenmeyer

December 12, 2001
Objective:

We want to predict the number of games that an NBA team wins (in an 82-game season) based on relevant season statistics such as field goal percentage, free throw percentage, three point percentage, rebounds-per-game, turnovers-per-game, and assists-per-game, as well as similar statistics for its opponents. We will select variables by using all subsets regression on a training set from the available data. After making any necessary transformations, we will fit that model to the test set and use the results for inference.

Data Description:

The data for our project was collected from the following website:

http://www.allsports.com/nba/stats/Stats/ns4stats.htm

This data was collected very simply; the team statistics for each game were recorded and then compiled at the end of the season on this comprehensive site of individual and team statistics for NBA players and teams. We collected only raw data from this website and then calculated the percentages and per-game averages in SAS.

The population from which we drew the data is all NBA seasons since its birth in 1949. Since our sample consists of only the last 12 full seasons, it is not a random sample from that population. However, this was an unavoidable consequence of not being able to find comprehensive stats for seasons prior to 1988. In fact, it is doubtful whether they were even collected during the first few decades of the NBA’s existence. Therefore, the application of any conclusions should be restricted to the current era of NBA competition.
**Data Analysis:**

**Methods:**

First, we split our data into two sets. The first, our training set, consisted of the nine seasons between 1988-1997, while the second, our test set, consisted of the three most recent full seasons (not shortened by strike). Thus, the training set took 75 percent of the data, while the test set consists of the other quarter.

We needed first to find a good model using our training set. In order to determine which variables would best predict wins, we used all subsets regressions to select variables. We entered field goal percentage, free throw percentage, three point percentage, rebounds-per-game, turnovers-per-game, assists-per-game, and personal fouls-per-game for each team, along with the corresponding statistics for its opponents, as the candidate variables for selection.

From the graph of $R^2$ (p. G1), we decided that approximately six variables should be included in the model, since the best model with seven predictors improved the $R^2$ value by less than .02. However, we examined four models, the best model with five predictors, the best with seven predictors, and the two best with six predictors (p. 1).

We examined both the residual and the partial regression plots for each model, and finally settled on the model with six predictors that yielded the highest $R^2$. This model includes the following six predictors: field goal percentage, rebounds-per-game, turnovers-per-game, opponents’ field goal percentage, opponents’ rebounds-per-game, opponents’ turnovers-per-game. The residuals are evenly scattered about zero with no systematic deviation or curvature (p. G2).
The variance appears to be constant as well. Also, the partial regression plots are all linear (pp. 4-9), and the normal probability plot is very straight (p. G3). (These plots are so good that we can hardly believe it!) Since the residual and partial regression plots are more than satisfactory, no transformations are needed. Thus, we decided that this model was the best because it is fairly simple and fits the data well.

With our model chosen, we turned to our test set. Out of curiosity, we ran all subsets regression again on this set, using the same candidate variables for selection. Interestingly enough, we found, using the .02 rule-of-thumb for $R^2$ (p. G4), that our best model in this case actually included only four predictors – field goal percentage, turnovers-per-game, opponents’ field goal percentage, and opponents’ turnovers-per-game – and that these four variables were a subset of the six variables included in the best model we obtained from our training set (p. 10).

Since the best model chosen from the test set includes only four variables and achieves a higher $R^2$ than the six-variable model chosen from the training set, it is possible that the variance of wins is smaller for these three most recent seasons (seasons used in the test set). To investigate this idea, we joined the training and test sets and ordered the data by season from earliest to latest. We then plotted the residuals against observation number to examine any changes in the variance over time. However, the variance appears to be constant over time (p. G7).
Interestingly enough, we found that the model with six predictors that yielded the highest $R^2$ (using our test set) was the same as the best model from our training set (p. 10). Therefore, we fit this six-predictor model using the test set.

Results:

As before, the residuals appear to have mean zero and constant variance (p. G5). The partial regression plots are very linear (pp. 13-18) and the normal probability plot is very straight (p. G6). The fitted regression equation of the model on the test set is:

\[
WINS = 21.0165 + 418.159\text{FGPER} + 1.268\text{REBPERGAME} - 4.472\text{TOPERGAME} - 349.146\text{OFGPER} - 1.026\text{OREBPERGAME} + 3.029\text{OTOPERGAME}
\]

When fit to the training set, this model yielded an $R^2$ value of .8562, but when it was fit to the test set, the $R^2$ was even higher at .9076. From the SAS output (p. 11), we can see that all the predictors are significant, with all p-values much less than .05.

Also, multicollinearity is not a problem, as all the predictors have variance inflation factors between 1 and 2 (p. 11). Thus, if one or more of the variables were removed from the model, the regression coefficients for the remaining predictors would not change significantly.

High leverage points would have leverage greater than $(2p/n) = (14/87) = .1609$. Observations 2, 17, and 33 are high leverage points but are not cause for much concern (p. 21).
Conclusions:

Through our variable selection, we determined that the most important predictors of a team’s wins are its field goal percentage, rebounds-per-game, and turnovers-per-game, along with the same statistics for its opponents. While this may seem like common sense to the knowledgeable basketball fan, *how much* a particular statistic affects a team’s wins is very interesting. For example, if all other statistics are held constant and a team increases its field goal percentage by one point, we can be 95% confident that the team will improve by an average (rough) of 3.5 to 5 wins (from the confidence intervals for the regression coefficients). Also, if a team grabbed an extra rebound each game (and the other statistics remained constant), it would earn an average (again rough) of 0.5 to 2 more wins during the season (p. 11).

Our regression equation is reasonable in terms of the signs of the regression coefficients. To be specific, a higher field goal percentage and more rebounds-per-game should increase a team’s wins (positive signs), while accurate shooting and aggressive rebounding by the opposition should decrease a team’s wins (negative signs). Moreover, turnovers will hinder a team (negative sign), whereas opponents’ turnovers contribute to success (positive sign).

We can make up statistics for a “fantasy” team and determine, using our equation, how many wins that team would earn over the course of a season. If a team made 45.5% of its shots, grabbed 43.9 rebounds-per-game, and had 13.6 turnovers-per-game, and its opponents’ posted a field-goal percentage of .433, pulled down 41.1 rebounds-per-game, and committed 15.1 turnovers-per-game,
we can predict with 90% confidence that this team will earn between 51 and 66 wins (p. 20). This is a legitimate prediction (not an extrapolation) since the leverage for this combination of predictor values is .0507 (p. 21).

Further analysis could be done at a different level of basketball competition (college or high school) to determine if the same variables are significant predictors of wins or if other variables prove more important. This would most likely require a prediction of winning percentage instead of wins (because not all college/high school teams play an equal number of games) but that would not be a difficult adjustment.

Also, we could further test the accuracy and validity of our regression equation during the current NBA. We could use our equation and each team’s mid-season statistics to predict their mid-season wins, and then compare our predicted values to the actual observed values.