Quantifying Persistence in Run Model Errors Across Team-Seasons

by Theodore L. Turocy

Department of Economics
Texas A&M University
College Station TX 77843

February 7, 2004

The baseball analysis community has estimated many different specifications for "production functions" for runs scored, both linear and nonlinear, over the last few decades. In general, it has been observed that teams that overperform or underperform relative to their predicted run production in a season will tend to regress to the mean in the following season. This suggests that there are generally no important team-specific features which persist from season to season that are not captured as inputs into these functions. This note seeks to quantitatively verify this folk wisdom in the context of specifications estimated by linear regression.

I use team-season-level data from 1974 through 1992, inclusive, and denominate all quantities in per-game-played terms. I investigate two specifications. The first specification is a basic linear estimator, where the vector of regressors is (1B,2B,3B,HR,BB,HP). The second specification is an extended specification designed in Turocy (2003)[1] to correctly account for team speed in the estimation; the estimated specification is called Model 4 in that paper. Essentially, this specification adds data on stolen bases and caught stealing (separated by attempts of second base and third base), grounded into double plays, and advancement on errors.

The ordinary least squares fit of the basic model to team offensive data gives a standard error of .1484 runs per game, and an autocorrelation of the residuals of .1204. There are 488 team-seasons in the sample; simulation of a like number of independent normal random variables with standard deviation .1484 resulted in an empirical autocorrelation of the residuals greater than .1204 in only 4 out of 1000 simulations, implying that the autocorrelation is significant. The full model fit to the offensive data drops the standard error to .1382, and gives residuals with an autocorrelation of .0598; in 1000 simulations, the autocorrelation exceeded .0598 83 times, putting this autocorrelation near the boundary of significance at the standard levels.

This document has been produced using GNU $\TeX_{\text{MACS}}$ (see http://www.texmacs.org).

1. "Offensive Performance, Omitted Variables, and the Value of Speed in Baseball." Available online at http://econweb.tamu.edu/turocy/papers/runest.htm

The difference between the specifications is the inclusion of factors related to speed in the full model. To see if this is in fact what is being controlled, the same specifications are estimated using data for teams on defense. In this case, the basic specification results in a standard error of .1499 with an autocorrelation of .0932 (simulated p-value .027) and the full specification a standard error of .1445 (simulated p-value .067). Inclusion of the additional regressors does not reduce the standard error and observed autocorrelation of residuals as much as when using offensive data, suggesting that these additional regressors are in fact picking up primarily differences in speed.

To correctly account for this autocorrelation, I assume that the error term in the regression, instead of being uncorrelated, follows a first-order autoregressive process given by $\epsilon_{ti} = \rho\epsilon_{t,i-1} + u_{ti}$, where u_{ti} is independent across teams and seasons (where i indexes teams and t indexes time). The value of ρ is estimated by the Cochrane-Orcutt procedure, which chooses ρ such that the sum of squares of residuals of the corresponding ordinary-least-squares estimators is minimized. The resulting estimates of ρ closely correspond to the autocorrelations observed above.

Specification	Estimated p
Basic, offense	.124
Full, offense	.063
Basic, defense	.096
Full, defense	.076

Importantly, however, the coefficient estimates obtained from the regression model augmented with autocorrelated errors do not differ in any significant fashion from those in the regressions ignoring the autocorrelation. Therefore, neglecting the autocorrelation in estimating the parameters of these type of models is likely not a problem.

Even though the autocorrelation does appear to be statistically significant, it is not substantial in magnitude. The autocorrelation of about .12 in the basic offense model implies that a team that outperforms the model's prediction by 30 runs in a season would expect, on average, to outperform the model's prediction by only about 4 runs the following campaign.

An interesting question is whether adding further factors to the model could further reduce the autocorrelation. Some candidates for sources of the autocorrelation of the residuals include strategy and managing styles, and team composition (balanced versus unbalanced lineups, for example). Also, note that while this estimates a linear approximation to the production of runs, the true function is nonlinear, but with a modest curvature. This alone could account for the autocorrelation, since teams that score many runs in one season are likely to do so in the next as well (the autocorrelation of runs per game in the sample is .42).