Sports Won-Loss Records Correlations

(The attached PDF file has better formatting.)

{Section 4 of the original paper has additional background. We examine the correlations here, not the χ-squared test. This part is *not required* for the student project.}

Take heed: The student project does not require you to read the *PCAS* paper. Complete instructions are posted on the discussion forum, along with illustrative worksheets. Past student project using this project template are posted on the discussion forum, so you can see what others have done. But some items are better explained in the original paper. This paper is on the CAS Exam 9 syllabus. If you take CAS exams, you read this paper anyway.

Take heed: If you have taken the time series course, you can form the correlogram of the won-loss records. This examines the same relation as the correlations here. It is *not necessary for the regression analysis project*, but you will find it useful. It shows the correlations graphically and helps you select the optimal regression equation.

Step #1: Determine if past won-loss records are a valid predictor of future winnings. Order the teams any way you like and compute their winning percentage in two adjoining years.

Illustration: Suppose we have ten teams, which we label T_0, T_2, T_3, …, T_{10}. Form two series: won-loss records for 2004 and for 2005. Form the correlation between these two series.

~ If the correlation is positive, teams which did better in 2004 also do better in 2005.
~ If the correlation is negative, teams which did better in 2004 do worse in 2005.

You can do this for several pairs of years, such as 2001 with 2002, 2002 with 2003, 2003 with 2004, and 2004 with 2005, and take an average.

If the correlation is not significantly different from zero, check your work. If you have not mis-labeled the teams, choose a different sport, league, country, or years. For the major league U.S. sports, the correlation should be positive.

Take heed: The illustrative worksheet on the discussion forum shows the computations. A separate discussion forum posting documents the illustrative worksheet.

Step #2: Determine if older won-loss records are a less useful predictor of future winnings. Do the previous exercise with a longer lag between the years. For a 2 year lag, correlate 2000 with 2002, 2001 with 2003, and so forth.

The correlation declines as the lag increases. For baseball, the correlation declines to zero within ten years. For some others sports, the decline is more rapid.

The illustrative worksheets code the cell formulas for these correlations. The student project does not test your Excel expertise.

CORRELATION WORKSHEET

{This dialogue explains the correlations among years. The correlations help you interpret the *t* statistics when the explanatory variables are correlated.

- Most candidates do not need explanations of Excel's autofill feature or instructions on cutting, pasting, and deleting.
- The detailed instructions are for candidates who rarely use Excel. Skip any sections you do not need.

You don't have to use correlations for a student project on won-loss records. But the correlations help you understand the relations, so we explain them.}

The illustrative worksheets show *correlations*, *correlograms*, *regressions*, and *F* tests. Correlograms are taught in the time series on-line course. They are not needed for the student project, but candidates taking the time series course may want to calculate them.

Jacob: What do the correlations show?

Rachel: The R^2 of the regression equation measures how well the independent variables explain the dependent variable. The worksheet for correlations shows another perspective on the relation of these variables.

- For a two-variable regression model, the R^2 is the square of the correlation.
- For a multiple regression model, the correlations are the square root of the R^2 *if each independent variable is used alone*. Instead of the correlations, we may compute the R^2 of a set of simple linear regression equations.

Jacob: The *t* statistics measure whether each variable is significant. Do the correlations add anything to what we learn from the *t* statistics of the multiple regression equation?

Rachel: When independent variables are correlated, the multicollinearity distorts the β parameters. The multicollinearity can be severe in a regression using lagged variables.

Take heed: The time series on-line course covers this topic. For the regression analysis student project, know that the multicollinearity affects the β estimates.

Jacob: How can we see this multicollinearity?

Rachel: We see the multicollinearity several ways:

- As we add independent variables, the standard error of the regression decreases. If the independent variables are orthogonal (independent, uncorrelated), the standard errors of the β parameters also decrease. If the independent variables are correlated, the standard errors of the β parameters may increase. In this student project, the won-loss records by year are correlated. As we use more years, the standard errors of the β parameters increase, and the results of hypothesis testing become unstable.

- Any independent variable used alone may have a large β and a high *t* statistic. When the independent variables are used in combination and they are highly correlated, the β parameters are smaller and the *t* statistics are lower.
- The first several correlations are positive. The multiple regression equations may have negative β coefficients in later years because of the multicollinearity. The negative ß's and their *t* statistics do not indicate the predictive value of those independent variables. If you obtain negative ß's, you may be using too many years.

Take heed: The comments and call-outs in the illustrative worksheets highlight these items.

Jacob: How do we calculate the correlations?

Rachel: We form an array of won-loss records by team and year. The illustrative work-sheet shows how to form the correlations for the years 1901-1960. The 8 teams in a League are the columns of the array and the 60 years 1901-1960 are the rows.

Take heed: Your student project may use other teams and years (or another sport). You can copy sections of the illustrative work-sheet, adjust the parameters for the number of teams and years, and compute the correlations.

In the second part of the student project, we compare two sets of teams, such as National vs American League teams, using an *F* test. Look at the correlations in the two Leagues: they are close but not identical. The statistical analysis says whether the differences can be attributed to random fluctuations.

Your student project may use a different *F* test. You may compare two time periods, or you may compare two types of teams (such as good teams vs bad teams). The *F* statistic is complex, and it is hard to know if you are doing the procedure correctly. Comparing the correlation matrices for the two time periods or the two types of teams tells you whether to expect similar regression equations.

On the illustrative worksheet labeled *American correlations*, the teams are in Columns C through J. The won-loss records are in Row 15 for 1960, Row 16 for 1959, Row 17 for 1958, and so forth.

Jacob: Do we use the same procedure for correlations in 1961 – 2005?

Rachel: The teams and the number of games in the baseball season change in 1961. (The 162 game season began in 1961.) Missing data can distort the regression analysis. If you use four past years and a team did not exist in 1990, don't include that team in the regression analysis until 1995.

For the student project, select teams and years. Avoid teams that start mid-way through the experience period, unless you explicitly compare new vs established teams.

Take heed: Using a subset of teams may cause the *average* won-loss record to differ from 50%. Ideally, you normalize the data, but this is not required for the student project. The illustrative worksheets compare Boston vs New York both wit and without normalization. The project template explains how normalization affects the regression equations and the *F* test. If your student project compares data sets with materially different, you may want to normalize the won-loss records.

Jacob: Where are the correlations?

Rachel: The correlations are to the right side of this matrix. The correlations of lag 1 are in Column L; the correlations of lag 2 are in Column M; and so forth.

Jacob: What is a correlation of lag 1?

Rachel: The correlations of lag 1 are for Years Z and Z+1. We show this correlation in the row for year Z. We use Excel's *CORREL* built-in function. For the correlation in the 1959 row, we correlate the row of 8 losing percentages for 1959 with the row of 8 losing percentages for 1960. The formula in cell L16 is =CORREL($C16:$J16,$C15:$J15).

Jacob: Why do we use the mixed references: absolute references for columns and relative references for rows?

Rachel: The mixed references may it easier to copy across columns and down rows.

- When copying across columns for lags 2, 3, …, the column references do not change. We make one change in a row reference. (You can automate this as well. These instructions are for candidates with limited Excel knowledge.)
- When copying down rows, the row references change properly.

Jacob: Some values are missing in the correlations; why is that? For instance, the 1960 row has no entry for any lag; the 1959 column has no entry for lags 2 and greater.

Rachel: The first correlation of lag *k* appears in the row for 1960 – *k*. For example, the first correlation for lag 3 appears in the row for 1960 – 3 = 1957. The final result is a rectangle with a triangle missing at the top.

Jacob: How do we use these correlations?

Rachel: We form averages for each column in Row 75.

- If the influences on the won-loss records change over time (e.g., draft rules or free agency rules change), the correlations may change.
- If we don't see large changes, we assume the average is the expected correlation.

The correlations tell us the *maximum* number of years to use in the regression analysis. If the correlation for lag *k* is significantly different from zero, we *might* use that lag in the regression analysis.

Jacob: Do we use the correlations like the correlogram in time series analysis?

Rachel: For time series, we examine the sample and partial autocorrelations to determine the maximum orders of the moving average and autoregressive parts of the ARIMA model. You don't need to compute correlograms or use ARIMA processes for this student project.

Jacob: What do we infer from the correlation matrix on the illustrative worksheet?

Rachel: Consider the correlations for American League teams from 1901 to 1960.

Lag	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*
Corr	*63.0%*	*50.6%*	*43.3%*	*37.2%*	*26.5%*	*21.4%*	*15.4%*	*12.1%*	*9.5%*	*10.2%*	*6.6%*	*6.6%*

The average correlation of lag 1 is 63.0%. If we regress a team's won-loss record in Year X on its won-loss record in Year X-1, the R^2 should be about $0.630^2 = 0.397 \approx 40\%$. It is useful to include Year X-1 in the regression equation.

The average correlation of lag 12 is 6.6%. If we regress a team's won-loss record in Year X on its won-loss record in Year X-12, the R^2 is about $0.066^2 = 0.004 \approx 0.4\%$. It is not useful to include Year X-12 in the regression equation.

Jacob: Even if the average correlation is only 6.6%, it is still positive. Shouldn't we include it in the regression equation?

Rachel: Suppose the optimal regression equation is $WLR_T = 5\% + 90\% \times WLR_{T-1} + \epsilon$. No other past years help the regression equation. This is a common type of formula; in the time series course, it is called an autoregressive model of order 1, or AR(1).

- The correlation of lag 2 is $90\%^2 = 81.00\%$
- The correlation of lag 3 is $90\%^3 = 72.90\%$
- …
- The correlation of lag 12 is $90\%^{12} = 28.24\%$

Even these high correlations do not mean that we should use the second and third past years in the regression equation. The 6.6% correlation for the twelfth prior year stems from the correlations for the first and second past years.

Jacob: Does the time series course say how many past years we should use?

Rachel: Choosing the optimal number of past years is a regression analysis issue. To fit an ARIMA model, we use multiple regression analysis.

Jacob: For the student project, how would we use these correlations?

Rachel: The correlations for the first two lags are significant. For the next four lags, the correlations are ambiguous. These past years *might* help the regression equation, but the improvement (if any) is small. Even for these four years, the adjusted R^2 may decrease and the β parameters may turn negative. For the next six years, the correlations are too small; we do not expect an improvement in the regression equation.

These correlations tell us what to expect. We do the full regression analysis, with the adjusted R^2, *t* statistics, and *F* statistics, for the student project.

Jacob: Should we do these correlations separately for the two leagues, or should we use all 16 teams?

Rachel: If we compare the two Leagues for the *F* ratio section of this student project, we examine the correlations separately by league.

{Pages 237-238 of the *PCAS* paper show the correlations. To check your results, compare your correlations with those in the *PCAS* paper.}

CORRELOGRAM WORKSHEET

{Jacob and Rachel discuss correlations vs correlograms and the sample autocorrelation function. Correlograms are in the time series course and are not needed for the student project on sport won-loss records. If you have not taken the time series course, ignore this section; it is not needed for the student project. The Excel code for the sample autocorrelation function is explained in the time series student project discussion forum.}

Jacob: Is the row of average correlations the same as the sample autocorrelation function?

Rachel: The row of average correlations is row 75 in the illustrative worksheet. The average correlation is *similar* to the sample autocorrelation, but it is not identical.

- The sample autocorrelation function measures the autocorrelation in a single series. The correlation of the Boston losing percentages from 1901-1959 with their losing percentages from 1902-1960 is the sample autocorrelation of lag 1. We show this on the *CORRELOGRAM* illustrative worksheet. The sample autocorrelation adjusts for degrees of freedom, so that we can compare sample autocorrelations of different lags. The sample autocorrelations are not needed for this student project; we show them for information only.
- The correlations here use the won-loss records of different teams in the same year. The correlation of lag k is the correlation of a series of 8 losing percentages in year T with 8 losing percentages in year T - k.

Jacob: Does the correlogram worksheet show the same information as the correlations worksheet?

Rachel: The correlogram worksheet shows the autocorrelations for a single team. The illustration uses Boston from the American League.

- The correlations worksheet correlates the losing percentages of the eight teams in the League between two years.
- The correlogram worksheet correlates the losing percentages of a single team with the losing percentages of the same team lagged k years.

The correlogram shows the decline of the sample autocorrelations.

- For the first 7 lags, the sample autocorrelations are 30% or more.
- For lags 8 and 9, the sample autocorrelations are the run-off from the previous lags.
- For lags 10 and higher, the autocorrelations are not significantly different from zero.

Jacob: Do the correlations worksheet and the correlogram worksheet support each other?

Rachel: These worksheets imply that we use at most 6 or 7 past years for the optimal regression equation.

Jacob: Why do you say *at most* 6 or 7 years? Why not exactly 6 or 7 years?

Rachel: Even if one past year is optimal, the sample autocorrelations might be significant for 6 or 7 years. Regression analysis shows the optimal number of past years.