PROJECT TEMPLATE ON REGRESSION ANALYSIS OF SPORTS WON-LOSS RECORDS

Updated: December 27, 2006

(The attached PDF file has better formatting.)

The project template for regression analysis of sports won-loss records forecasts the future winnings of sports teams from their historical experience. It is based on a seminal paper on baseball won-loss records published in the *Proceedings of the CAS* that is now on the CAS Exam 9 syllabus. The paper is delightfully written; dozens of past actuarial candidates have rated this paper one of the best readings on the CAS syllabus.

The student project replicates the analysis in this paper and extends it several ways. It focuses on the regression aspects of the analysis, not the actuarial items. The extensions use *F* statistics to test several additional hypotheses; they are not needed for the themes of the paper. You choose one additional hypothesis for the student project.

Original analysis: We forecast a team's winnings from its won-loss record in previous years. WLR_t is the team's won-loss record in year *t*.

$$WLR_t = \alpha + ß_1 \times WLR_{t-1} + ß_2 \times WLR_{t-1} + \ldots$$

Definition: The won-loss record is the percentage of games lost. (The original paper uses games lost in baseball as a metaphor for an insured's loss ratio.) For sports that allow ties, a tie is half a win and half a loss.

Intuitively, the β parameters should be positive: a team that did well in past years is likely to do well next year. The β parameters should decline as the years get older. Last year's won-loss record is a good indicator of a team's quality; the won-loss record from twenty years ago doesn't indicate much. We expect that $ß_1$ is positive and relatively high, such as 25% or 30%; $ß_2$ is positive and smaller, such as 15% or 20%; and so forth.

The original paper shows that the β parameters decline to zero within about ten years. In the student project, you will find that the β parameters are not significantly different from zero within four or five years.

The β parameters decline for two reasons.

~ The poorest performing teams get the highest draft picks. A team that does poorly one year is likely to do better the next year. The sports in this project template have drafts.
~ The teams' performance regresses toward the mean for many reasons: players get older, retire, are traded, or are injured.

But some teams remain consistently good or bad:

~ A good coach can improve a team's performance for several years.
~ Wealthy teams can afford high salaries, which may improve performance.

The regression equation is an autoregressive AR(p) model, which you may recognize from the time series course. We do *not* deal with the time series aspects of this regression for the student project. The original paper developed a formula for β_t based on the covariance of the won-loss records by year. This analysis is *not* used for the student project.

The original paper does not *explicitly* examine whether the regression equation is the same for all teams, all leagues, or all sports. The paper implicitly addresses this issue, since teams or leagues or sports that have different covariances among years have different regression equations.

The student project focuses on the statistical techniques from the on-line course, not on the actuarial (credibility) concepts in the original paper. It formulates a hypothesis, explains how we test it with the *F* ratio, determines the *F* statistic and the degrees of freedom, and explains whether we should reject the null hypothesis.

Data: The needed data is on the NEAS web site. You select data, put it in the form needed for the regression analysis, and use the Excel REGRESSION add-in to derive the regression coefficients. You don't have to use Excel; you can use any statistical software, such as SAS or Minitab or "R." If you use Excel, you can use built-in statistical functions or VBA for the student project. The REGRESSION add-in is the easiest, but it is not versatile.

You can use the data on the NEAS web site; it is sufficient for the student project. We provide the web sites containing the sports statistics. You can design a student project with other data from these web sites, as long as it applies the statistical techniques to actual data.

You can use other sports, countries, or leagues if you have the data. For the U.S., similar projects can be done for minor league baseball or college basketball, though it is harder to compile statistics. You can use data for soccer, basketball, or baseball from numerous other continents.

After reading the project template and looking at the sports statistics, you may think of other analyses that you prefer to do. Sports is statistics intensive. You may do a student project on batting averages, pitching records, points scored, or yards gained.

We encourage candidates to design their own projects. The *student project* in this posting refers to the project template here. If you design your own project, be sure to explain clearly any definitions or sports terms. Our statistics faculty are sports fans, so make your project clear.

STUDENT PROJECT: TWO PARTS

The student project has two parts. The first part solves for the optimal regression equation, using the basic regression analysis techniques:

~ Form correlations among years and explain their meaning.
~ Specify a linear relation using N past years.
~ Compute ordinary least squares estimators.
~ Check R^2, $\overline{R}^2$, and *t* statistics.
~ Select the optimal regression equation.

The second part of the student project asks you to choose a hypothesis that requires an *F* statistic to test. We give examples of such hypotheses below. The steps include

~ Formulate a testable hypothesis.
~ Determine the restricted and unrestricted equations.
~ Determine the degrees of freedom for the *F* statistic.
~ Compare the *F* statistic to the critical values and test the hypothesis.

The results depend on the sport, the teams, the years, and the hypothesis. The optimal regression equation might have two years and a high R^2 in one project and eight years with a low R^2 in another project.

If you follow this project template exactly, we check if

~ You compute the ordinary least squares estimators correctly.
~ You use the regression diagnostics properly to select the optimal regression equation
~ You set up the equations and use the *F* statistic properly.

You may discuss the statistical techniques on the discussion board, but use different data for the project itself. Don't copy results from the original paper:

~ The original paper used won-loss records for 1901 – 1960. Choose other years for your student project. We focus on the quality of the statistical reasoning, not the quantity of data. You need not use all years; 1976-2005 is sufficient.
~ The original paper used all teams. You can choose a subset of the teams. Some new teams do not have a full history. The original paper wanted a sample in which the team and the number of games did not change. We are less concerned with this.

Pick the sport you analyze: baseball, basketball, hockey, soccer, or football. Each sport differs, because of the number of players on the team, the expected sports life of players, the number of games in the season, the draft rules, and the free agent rules.

In basketball, a single first-round draft pick may transform a team from the worst in the league to the best. In baseball or football, a single draft pick has a smaller effect. The regression coefficients should be larger for baseball and football than basketball.

CORRELATIONS

Read Section 4 of the original paper, which is attached to this posting. This section is about baseball, with a bit of elementary statistics. We examine the correlations in the student project, not the χ-squared test.

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}. We form two series: their won-loss records for 2004 and for 2005. We 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, choose a different sport, league, country, or years. For the major league U.S. sports, the correlation should be positive.

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 REGRESSION EQUATION

We determine the optimal regression equation in two steps: (i) Given the number of past years (independent variables), we optimize the estimators, and (ii) We select the optimal number of past years based on the adjusted R^2, *t* statistics, and *F* statistics.

Keep your project manageable by limiting the number of teams and years. You might use 12 teams, a maximum of 10 independent variables for the regression equation, and experience from 1981 - 2005. The first year that we forecast is 1991, using 1981 - 1990 as the observed data. The last year that we forecast is 2005. We forecast a total of 15 years. This gives 12 teams × 15 years = 180 data points for the regression equation with 10 independent variables.

For the regression equation with 9 independent variables, we have 12 teams × 16 years = 192 data points. For each number of independent variables, we have a different number of data points.

The analysis sequence starts with 1 independent variable, and proceeds to more variables. Depending on the results, you may stop after 3, 4, or 5 independent variables.

If you are comfortable with Excel, the regressions are routine. Excel has automated all aspects of the regression analysis. You compile the regression results and select the optimal equation.

If you are not comfortable with Excel, choose a smaller data set. Use at least five past years for the regression equation and at least 80 data points. You have enough data for hundreds of data points, but you need not use all the points.

Use the following sequence:

Arrange the data in N rows by cutting and pasting, where N is the number of data points. If we have k independent variables (past years), each row has k+1 columns. For the 12 × 15 illustration above

~ The first 12 rows predict the 2005 won-loss record from the won-loss records in years 1995-2004 for 12 teams.
~ The next 12 rows predict the 2004 won-loss record from the won-loss records in years 1994-2003 for 12 teams.

Forecast Year	*Team*	*Dependent Variable*	*Independent Variables*				
		Year T	*Year T-1*	*Year T-2*	*Year T-3*	*Year T-4*	*Year T-5*
2005	NYY	65%	70%	69%			
	BRS	55%	75%	60%			
	CWS	60%	50%	40%			
	…						
2004	NYY						

A statistician's task includes compiling the data into the format needed for regression analysis. Each student project has different teams, years, sports, and hypotheses. Our faculty review if you have set up the data in a reasonable format for the regression analysis.

~ Any format that allows you to get regression results is reasonable.
~ If you format the data so that you can't perform the regression analysis, the format is not reasonable.

Depending on your choice for the second half of the student project, you may order the rows differently. If you compare National League teams vs American League teams, you may put all the National League teams first, followed by the American League teams.

You may find it easier to put all the years for Team #1 first, then all the years for Team #2, and so forth. The method of organizing the data depends on your versatility with Excel.

~ If you format the data by cut and paste, put the rows in the order of the teams.
~ If you read the data into a VBA array, the VBA macro does the regression analysis for whatever years are desired.

We provide raw data on the NEAS web site; you organize the data in the format needed for the regression analysis. The format depends on the hypothesis you test in the second part of the student project. You save time by thinking through your project before you format the data.

SEQUENCE OF REGRESSIONS

We do several regressions in sequence, using first one past year, then two past years, and so forth. We use the regression add-in to get R^2, adjusted R^2, and *t* statistics for each regression.

We use the $\overline{R}^2$ to select the optimal regression equation. Use the principle of parsimony. If the $\overline{R}^2$ is only slightly higher with an additional year, such as 50% for 3 years and 50.5% for 4 years, don't use the additional year. Simpler equations reduce the chances of making errors.

We don't need the *F* statistic to select the optimal regression equation. The *F* statistic shows greater significance with more years, but doesn't add anything to the $\overline{R}^2$.

We use the *F* statistic to test if the optimal regression equation is statistically significant. If the optimal equation has an *F* statistic that is not significant at the 10% level, check your work. It is possible that everything is correct, but it is more likely you have made an error.

You may find that the adjusted R^2 keeps increasing for all ten years. You will probably find that the *t* statistic for the oldest years are not significant. We advise you not to use more than seven years, since the *REGRESSION* add-in has a limit of 16 explanatory variables. If you use all ten years, you won't be able to easily form the *F* statistic in the second part of the student project.

You may find that using more years causes some β parameters to be negative, which seems counter-intuitive. The author of the original paper got negative β's for several past years. It is hard to judge if this result is real or spurious. If all the β parameters are positive for N years but one is negative when you use N+1 years, use N years.

You choose the number of years and the optimal β parameters. Choosing the number of years is partly subjective. Suppose we have an adjusted R^2 of 38% with four years and of 39% with five years, but the *t* statistic for the fifth year is not significant at the 90% level. Statisticians argue whether nine years or ten years is better. We recommend four years, which simplifies your student project.

You may normalize the regression parameters so that the mean won-loss record is 50%. If we use all teams, we get a mean won-loss record of 50%. If we use only some teams, we may get some other mean. This is an optional adjustment. If you are familiar with normalizing methods, you can use it; otherwise, leave it out.

Some candidates will set up the regression easily; others will find this more difficult. Discuss the project on the discussion forum so that you understand the objective, but submit your own project. We explain the Excel *REGRESSION* add-in with the project template for regression analysis in loss reserving. If you have trouble running the Excel functions, post a question on this discussion forum.

Some candidates worry: Did I get the right answer? There is no right answer. We examine if you performed the regression analysis correctly and if your choice of the regression equation is reasonable. The correlations depend on the sport, the teams, and the years.

PART 2: THE F TEST

The second part of the student project uses an *F* statistic to test a hypothesis. We segment the data into two or more groups, such as

~ National League vs American League (for baseball)
~ Better teams vs worse teams, based on the previous year's won-loss record
~ Baseball vs basketball (or hockey or football)
~ Won-loss records for particular teams (you may have several groups)
~ Years, such as pre- and post- the free agent rule in baseball

You may choose the groups in many ways; you are not restricted in your choice.

~ If you choose two sports, such as hockey vs football, the same regression equation is unlikely to be appropriate for both sports.
~ If you randomly select teams, such as teams east of the Mississippi River vs those west of the Mississippi River, the same regression equation is likely to be appropriate for both.

The student project shows that you understand how to use an *F* test. We are not concerned with the actual result.

For each scenario, the null hypothesis is that the same regression equation is appropriate for both segments. To simplify the analysis, we fix certain regression parameters. For example, we may fix the number of independent variables (past years) to five years. We

perform the analysis separately for each segment. We use an *F* test to determine if the same regression equation should be used for both segments.

If you find that the optimal regression equation uses six past years, not five past years, you can use either six or five past years in the regression equations. The student project checks if you understand how to apply an *F* test, not if you have found the optimal regression equations.

We caution against using more than seven past years, or eight explanatory variables (including the intercept). The *F* test doubles the number of explanatory variables. Excel's *REGRESSION* add-in allows a maximum of sixteen explanatory variables. Other software packages allow more variables, but if you use Excel, keep it simple.

Illustration: We compare the National and American Leagues. We fix the regression to five independent variables. Both Leagues have average won-loss records of 50%, so we need not normalize the won-loss records to 50%. We use an *F* test to see if the same regression equation should be used for both Leagues. If we use rich teams vs poor teams or good teams vs bad teams, the average won-loss records differ. To normalize the won-loss records, use deviations from the means, not absolute numbers.

For each scenario, the intuition differs. For National vs American Leagues, we have no reason to think the regression equations should differ. The objective of the student project is not to find a better way of forecasting baseball results but to apply the statistical techniques to real data. As you do the project:

~ State the null hypothesis (e.g., the two leagues have the same regression equation).
~ State the expected result if the null hypothesis is true. Explain the constrained and unconstrained regression equations. Explain the degrees of freedom for the *F* statistic and the distribution of the *F* statistic if the null hypothesis is true.
~ State the result you derived and the conclusion you drew. The conclusion should be a probabilistic statement: "If the null hypothesis is true, the probability of obtaining the observed results (or more extreme results) is Z%…"

The results may depend on the number of data points or years. We get more conclusive results with more data points. No general rule applies; the number of data points needed varies with the hypothesis. Use at least 80 data points, so that your results are significant.

{*Note:* We encourage you to design other projects. If you use other data, such as college football, and you have only 20 or 30 data points, that is fine. If you design an alternative project, don't restrict your design too much by the number of data points. But if you have only a dozen data points, you don't get reasonable results, so choose a different design.}

For each analysis, we use an appropriate null hypothesis. For example:

Among the better teams (based on the previous year's won-loss record), small differences in last year's won-loss record have little effect on the current year's won-loss record. For

the poorer quality teams, a worse won-loss record gives a higher draft pick. For these teams, β_1 may be low or even negative, since a worse won-loss record gives a higher draft pick and perhaps better performance the current year.

The effect of draft picks differs by sport. In basketball, with five starting players and only about eight who see much action, a single draft pick may turn a losing team into a winning team, so β_1 may be low or negative. For football, where 30+ players participate in each game, a single draft pick has less effect.

We may believe that the difference between the first draft pick and the second draft pick has a material effect on next year's won-loss record, but the difference between the 11th draft pick and the 12th draft pick does not have a material effect. This implies that good teams and bad teams have different regression coefficients.

Statisticians have analyzed these relations in some excellent studies, with results that seem surprising at first but make sense. One might think that a high draft pick should help a team become more profitable, because a star player increases attendance at games. But two items offset this relation:

~ Better players get higher salaries.
~ Many high draft picks are over paid. Lower draft picks are priced more accurately.

Statisticians use a variety of statistics to determine a player's contribution to the team's success. They relate the player's draft pick to his salary and his contribution to the team's success. A first draft pick costing $2 million a year is like two lower draft picks costing $1 million a year each.

A study of pro football concluded that high draft picks were so over-priced that their average effect on the team's financial performance is negative. The best ratio of performance to price was at about the 46th draft pick. This suggests that the entire first round of draft picks are poor additions to the team. If the team has a limited budget, the optimal strategy is to trade the first round draft pick for two or three lower draft picks or for slightly older players whose performance can be better estimated. The team's performance will improve.

We mention this to show the practical use of statistics. An actuary with a sound understanding of statistics might consult for professional sports teams to optimize trading strategies.}

Other factors affect won-loss records, such as the budget for players' salaries. High salary budgets in New York and Boston help keep baseball won-loss records high. We don't have a good public source for salary budgets, but many baseball fans know the relative sizes of these budgets. A higher budget may *raise the intercept* in the regression equation. If you find this effect for certain teams, and you can differentiate among the teams by salary budget, you can analyze this effect.

The free-agent rule makes salary budgets more important. If players can't choose among teams, salary budgets have less effect.

Write-Up

Submit a Word (or WordPerfect) document summarizing your project and an Excel file (or other file with data) showing the regression analysis and the *F* test. Many candidates find it easiest to place comments in their Excel file and submit that alone. It is difficult for our faculty to decipher what you have done from these comments alone. If you submit just an Excel file with no write-up, our faculty may return the student project and ask for a write-up.

If you don't use Word or WordPerfect, any text file is fine. The text should reference the analysis in the Excel file, but you should copy over the results into the text file. The text file should explain your null hypotheses and the conclusions you came to.

Further Documentation

We take great care not to over-specify the work required for the student project. The SOA wants to see independent student projects, and we do not give you a complete template in which you place sports statistics. Most of this project is organizing the data and forming the proper null hypotheses.

We as review the postings on the discussion board and the projects that are submitted, we will add comments on the discussion board explaining items that seem confusing. We will post some sample projects with comments by the NEAS faculty so that you see what is expected. You may review the original paper to see how the analysis was once done. Review the textbook sections on the *F* statistic.

Sports Statistics

Sports statistics are publicly available on dozens of web sites. We provide files of won-loss records on the NEAS web site, which contain the data for this student project. You may find better data sources on other web sites, and you should feel free to list these sources on the discussion board or in your student project.