PROJECT TEMPLATE ON REGRESSION ANALYSIS OF SPORTS WON-LOSS RECORDS

(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 past experience. It is based on a 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 other hypotheses.
- You choose one or more additional hypotheses for the student project.

We provide illustrative worksheets and PDF files explaining the student project.

- The Excel worksheets compare Boston and New York baseball teams.
- The PDF files document the worksheets and explain the procedures in the 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:

- β's should be positive: a team that did well in the past is likely to do well next year.
- β's 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.

Illustration: 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 ten years.
- Your student project will probably show β's declining to zero within 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.

Take heed: 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

For student projects that you design yourself, you compile data from internet sites or other sources. The project template on sports scores provides 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." These packages have some features that Excel does not have.
- If you use Excel, you can use built-in statistical functions or VBA.
- The REGRESSION add-in is the easiest, and it is sufficient for the student project.

The data on the NEAS web site is sufficient for the student project. We provide web sites containing the sports statistics. You can design a student project with other data from these sites or other web sites, as long as it applies 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 get statistics. You can use data for soccer, basketball, or baseball from other countries.

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 you to design your own project. The *student project* in this posting refers to the project template here. If you design your own project, explain any definitions or sports terms. Our statistics faculty are *not* 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 uses an *F* statistic to test a hypothesis. We give examples of such hypotheses below. The steps are

~ Formulate a testable hypothesis and select a critical value.
~ 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 five years with a low R^2 in another project.

If you follow this project template, we check how

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

You may discuss the statistical techniques on the discussion board, but use different data for the project. Don't copy results from the illustrative worksheet on the discussion forum

~ 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.

Illustration: In basketball, a 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.

Take heed: The illustrative worksheets are explained in separate PDF files: correlations, regressions, and *F* test. Read the PDF files, look through the illustrative worksheets, and check the data on the NEAS web site (and other sports sites if you wish). If you like sports, this student project is wonderful way to fulfill VEE requirements.

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.

- Some candidates place comments in the Excel file and submit that alone.
- Our faculty can not decipher what you have done from these comments alone.

If you submit just an Excel file with no write-up, the course instructor may return the student project and ask for a write-up.

You can use Word, WordPerfect, Adobe Acrobat, Notepad, or any text file. The text may reference the analysis in the Excel file, but you should copy important results into the text file. The text file should explain your null hypotheses and the conclusions you came to.

Do not write: “The statistical analysis is documented in the Excel file.” The SOA wants evidence that you understand what you are doing. The student project is not graded on writing style, but it must show that you understand the statistical procedures.