## SPORTS SCORES REGRESSION IDEAS

(The attached PDF file has better formatting.)

This project template suggests several hypotheses for your student project. You can find other hypotheses several ways.

- Other NEAS postings on the discussion forum have more suggestions.
- The past student projects on the discussion forum show what others have done.
- Candidates may post topics on the discussion forum.

Hearing the hypotheses used by other candidates gives you ideas for your student project. Post your hypothesis and conclusion on the discussion forum.

*Illustration:* "I compared the two divisions in Sport XYZ for years 1975 - 2005. My optimal regression equation had four past years for the Eastern Division and three past years for the Western Division. I used three past years (or four past years) for the *F* test. The result was significant (or not significant) at the 5% level, or the *F* test had a *p* value of 7.8%."

Some hypotheses are different. You might write: "For each year, I chose the team with the best won-loss record and the team with the worst won-loss record in each League, which I labeled good and bad teams. My hypothesis is that good teams have a high $\beta_1$ coefficient, since they have a low draft pick, and bad teams have a low $\beta_1$ coefficient, since they have a high draft pick. The coefficients in the optimal regression equations are … An F test shows this difference is significant at the 1% level."

If you use another sport or years, your posting is especially valuable to other candidates. Comparing the optimal number of years in each sport is a good student project.

*Illustration:* Your student project might say: "I estimated the $\beta$ of the past year for four sports: baseball, basketball, hockey, and football. I then regressed the $\beta$ on the number of games in a season and the number of players on a team. My hypothesis is that both explanatory variables should have positive coefficients (give reasons). I found that …"

The optimal number of years may change over time. You may compare old years (1901-1960) with new years (1961-2005). If the difference is significant, you may suggest reasons why it occurs.

You may compare Super Bowl winners (or World Series Winners or NBA champions) in each year vs other teams. You can post your hypothesis along with the list of teams and years and your results.

You can use statistics besides won-loss records. Not all candidates realize the wealth of statistics data on the internet.

*Illustration:* Your write-up might say: “My student project examines if past batting averages are a good predictor of future batting averages. I use an *F* test to determine if the same regression equation is valid for young vs old players. I assumed that younger players have increasing batting averages and older players have decreasing batting averages. I used players who have already retired, and I normalized each player’s lifetime batting average to one. I fitted regression models to young vs old players.…

“I defined young players as those under 30 and old players as those over 29 …” <OR> “My initial analysis showed that batting averages are best predicted from batting averages of the past three years. I divided each player’s N-year career into three parts:

- The first three years, which are not predicted.
- The next ½ × (N – 3) years = young players.
- The last ½ × (N – 3) years = old players.
- If N – 3 is odd, the middle year is defined as an old player.

“Batting averages differ by player and between young vs old years for the same player. I normalized each player’s batting average to 100%, giving equal weight to each year.”

“The optimal regression equations are … for all players, … for young players, and … for old players. I used an *F* test to see if the same regression equation might be used for all players.…” (Your write-up would show the degrees of freedom and number of restrictions; the $R^2$ or the ESS (or RSS) of the restricted and unrestricted equations; and the *F* statistic and either the *p* value or the critical values at important significance levels.}

“The *p* value for the *F* test is 14.2%. This value is not highly significant, but the observed difference is as expected: younger players have lower β coefficients and increasing batting averages. I conclude that young and old players have different regression equations.”

You may discuss how best to differentiate good from bad teams.

*Illustration:* “I compare good vs bad teams, using three definitions.

A. Teams with ten year won-loss records above or below 50%.
B. Teams with won-loss records above or below 50% each year.
C. The best and worst teams each year.

My *F* test shows

- No significant difference for Definition A (*p* value = 60%).
- No conclusive result for Definition B (*p* value = 14%).
- A significant difference for Definition C (*p* value = 3%).

*INTERNET SURFING*

If you want something different, surf the web. You may want a student project on your home city's team, but using a different statistic that those in this project template.

The team probably has a web site, which has years of sports statistics.  Visit the site, and imagine the potential relations among the statistics.