SPORTS SCORES STUDENT PROJECT STEP BY STEP GUIDE

(The attached PDF file has better formatting.)

Jacob: Should the student project be like the analysis in the illustrative worksheets? How closely should we follow the project template?

Rachel: The project templates and illustrative worksheets are samples. They give ideas; they are not formulas for your student project.

You choose the analyses in the student project. Many student projects are unrelated to the project templates. Pick a topic, formulate a hypothesis, use statistical techniques to test the hypothesis, write up the results, and send in the student project.

- If you design your own student project, we give credit for reasonable analysis applying statistical techniques to real data. Check if the results are reasonable; if not, review the project template to be sure your work is correct. But do not worry If your analysis does not support the hypothesis, explain what you expected and what the data show.
- If you follow a project template, your student project should differ from the illustrative spread-sheets. We want to see that you understand what you are doing.

Jacob: I want to do a student project on sports statistics, but I'm not sure what to select. Can you give a step-by-step guide to each part of the student project?

Rachel: Use the following steps:

Step #1: Select a sport, a time period, and sets of teams. The *PCAS* paper uses major league baseball for 1901-1960 and sixteen teams. The illustrative worksheets show Excel formulas for the common techniques, using Boston and New York won-loss records.

Data for major league baseball, basketball, hockey, and football are on the NEAS web site. If you use minor league, college, or non-U.S. games, or if you examine other relations, you have to compile data. Numerous web sites have sports data.

Don't just replicate the illustrative worksheets, but you don't have to change everything. You may use more recent years, another sport, another statistic, or a different *F* test.

We explain the *F* statistic in detail. We use an *F* test to compare Boston and New York.

Take heed: If you are uncertain about the method, pick two teams and replicate the Boston vs New York analysis on the illustrative worksheets. Spend an hour working through the work-sheets. Once you understand the procedures, select the data for your student project.

Take heed: Some candidates seek assurance that their proposed project meets the VEE requirements. We can not determine this until we see the completed project.

- Designing a different project demonstrates competence with the statistical techniques, and we require less rigorous execution.
- Pick a topic you like, and write a student project. Don't worry that the topic is not good enough or the results are not adequate.
- We do not judge if the results are correct. We review your project to see if you understand the statistical techniques.

Step #2: Posit a relation to examine. The illustrative worksheets examine whether past won-loss records predict future won-loss records. Other relations might be

~ Player salaries affect the won-loss record.
~ Batting percentages or home run totals affect won-loss records.
~ Pitching statistics predict won-loss records.
~ Points scored per game (basketball) predict won-loss records.

Each sport has statistics that you can use, and many teams have extensive web sites.

- Choose a sport and surf the web sites of several teams. Much information is on private web sites run by sports fans and freely available to others.
- You don't need a hypothesis before you start. The data on sports web sites give you ideas for a student project.

Illustration: The relations of various statistics to won-loss records is not always clear. You can regress won-loss records on home-run totals, batting averages, strike-outs, errors, stolen bases, and other items.

Step #3: State the null hypothesis.

- The first part of the project template tests if past won-loss records predict future performance. The null hypothesis is that they do not.
- The second part of the project template tests if the regression equations differ for two sets of teams. The null hypothesis is that they do not.

Illustration: The first part of your student project might relate won-loss records to batting averages and home-run totals. The second part of your student project might test if this relation differs for 1901-1960 vs 1961-2005.

If possible, do preliminary tests of the hypothesis. The illustrative worksheets examine the correlations between years. If the correlation of lag 1 is not significantly different from zero, past experience does not affect future experience.

Take heed: If your preliminary analysis shows no significant relation, choose another null hypothesis or other data. You need a hypothesis for which you can perform an *F* test.

Illustration: Don't use red vs blue teams in Camp Color-war. One year's results have no effect on the next year's results. Similarly, don't use the score in the all-star game or the world series games. We don't expect any predictive value for next year's score. (In truth, the All-Star games results seem not to be a random draw with a 50% probability, but the probability changes from year to year. The All Star game results are binary (win or lose), and they are not amenable to the techniques taught in the on-line courses.)

Step #4: Set up the regression equation to test the null hypothesis. The student project focuses on optimizing regression equations and *F* tests.

Illustration: An ideal student project has a non-trivial result. The illustrative worksheets compare Boston and New York. The material difference in team quality for 1901-1960 suggests that the teams have different regression equations. After normalization, the *F* test shows almost identical equations.

Step #5: Optimize the regression equation. Determine if it is best to use 1 year, 2 years, …, to predict future performance.

- The optimum is not always clear. The R^2 always increases with the number of years.
- Use the adjusted R^2, *t* statistics, the *F* statistic, intuition, and parsimony.

Your write-up explains the factors you use. There is no single correct answer. Even for a given data set, different statisticians may not agree on the optimal regression equation.

Illustration: Suppose four past years gives β coefficients of 20%, 15%, 10%, and 5% for an adjusted R^2 of 36%. Five past years gives β coefficients of 20%, 16%, 14%, –5%, and 5% for an adjusted R^2 of 38%.

- The 2 percentage point increase in the adjusted R^2 is material.
- But one might suspect that the five year coefficients are distorted by multicollinearity among years and statistical fluctuations in the data.

The choice of the optimal regression equation would depend on significance of the –5% coefficient in the five year data and the intuitive justification for the negative coefficient.

Some candidates want exact guidance. They ask: "What significance level should we use?" and "What does intuition mean?" The student project shows that you can apply statistical techniques to real data, not that you can follow instructions.

~ Choose a significance level and explain why it is appropriate.
~ Intuition implies judgment. If we prescribe everything, nothing is left for your judgment.

Step #6: Understand the Data

The regression equation depends on the consistency of the data. If the draft rules or free agency rules change, the optimal regression equation changes. You are not expected to know these rules. This is a statistics course, not a sports course (so you may ignore this step), but statisticians who understand their subjects do better analyses.

We have listed references to web sites with the current rules. You may discuss on the discussion forum items that might affect the regression. A student project is not graded adversely because it fails to consider a sports rule change.

The student project is not restricted to U.S. teams. The illustrative worksheets use U.S. teams because the data are available. If you live in Asia, you can use Asian sports teams.

Step #7: Choose a hypothesis to examine with an *F* test. The discussion forum postings explains how to compare two teams or two leagues (conferences, divisions). We suggest several other null hypotheses. Don't hesitate to use your own ideas.

Step #8: Begin with your optimal regression equation. If it has more than 8 independent variables, you have more than 16 independent variables in the unconstrained regression equation for the *F* statistic, which is more than the Excel *REGRESSION* add-in will handle.

Even 16 variables is cumbersome. We recommend a maximum of five past years for the *F* statistic. Few sports relations use more than five or six explanatory variables. If you get a complex regression equation, you may be making an error.

Five years is the recommended maximum, not the expected number of years. The illustrative worksheet shows an optimum of two years for the Boston baseball team and one past year for New York. A regression with one, two, or three past years is fine.

Step #9: Set up the constrained and unconstrained regression equations. The dummy variable may differentiate among Leagues, divisions, years, quality of teams, wealth of teams, or other items. You may also use separate regression equations.

Review the course textbook (or Mahler's *Guide to Regression*) on the *F* test and dummy variables. Excel gives you all the input figures, including the critical values.

Step #10: Derive R^2, RSS, ESS, and TSS for the constrained and unconstrained equations. Not all these are needed; explain which you use. Determine the degrees of freedom. Form the *F* statistic and explain what it shows. The textbook explains these items; review the sections on the *F* statistic to be sure your technique is correct.

State the conclusion of the test in probabilistic form. Excel (or any other statistical package) gives exact *p*-values. State the *p* value and explain its implications.

YOUR STUDENT PROJECT

Jacob: Some instructions say to design an independent student project. Other instructions explain how to replicate the illustrative worksheets. What is expected from us?

Rachel: The project template optimizes your efficiency. Begin by replicating the analysis in the illustrative worksheets with other teams. Once you understand the procedure, design your own project, using other teams, sports, years, and null hypothesis for the *F* test.

We have included enough guidance that you can complete the project and many ideas for null hypotheses that you can test. Many candidates find that statistical concepts that were fuzzy at first became clear from the student project.