SPORTS SCORES PROJECT TEMPLATE REGRESSION EQUATION

(The attached PDF file has better formatting.)

{This document explains the regression equations in the project template on sports won-loss records, documents the illustrative worksheets, and suggests topics for your project.}

Determine the optimal regression equation in two steps:

- Given the number of past years (independent variables), optimize the estimators.
- Select the optimal number of past years based on the adjusted R^2 and *t* statistics.

Take heed: 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 for 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 regression equations 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.

Begin with one independent variable, and proceed to more variables. Depending on the results, you may stop after 3, 4, or 5 independent variables.

Take heed: Use Excel's built-in functions and add-ins. You compile the regression results and select the optimal equation. 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						

Each student project has different teams, years, sports, and hypotheses. Part of the student project is setting up the data in a reasonable format.

~ Any format that allows you to get regression results is reasonable.
~ If you can't perform the regression analysis easily, either the format is not reasonable or you are not using the *REGRESSION* add-in correctly.

Depending on your hypothesis (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.

Summary: 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.

SEQUENCE OF REGRESSIONS

- Do the regressions in sequence: first one past year, then two past years, and so forth.
- Use the regression add-in to get R^2, adjusted R^2, and *t* statistics for each regression.

Use the $\overline{R}^2$ and the principle of parsimony to select the optimal regression equation. 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.

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.

It is unlikely that the adjusted R^2 increases for all ten years. Even if it does, you will find that the *t* statistic for the oldest years are not significant. Do not use more than seven years, since the *REGRESSION* add-in has a limit of 16 explanatory variables. If you use ten years, you can't 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 four or five 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, use them; otherwise, leave them out.*

Some candidates will set up the regression easily; others 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.

REGRESSION WORKSHEETS: BOSTON AND NEW YORK 1901-1960

{This dialogue explains the regression analysis in the illustrative worksheet. You format the data for the regression add-in and compare the output for each set of past years.

- The dialogue explains how to re-format the data.
- If you are proficient with Excel, you may not need the guidance here.

Student projects on the *F* test have several steps. The project template uses sports data, since many candidates enjoy a project about their home team.

- The illustrative worksheets compare Boston and New York.
- Focus on the concepts, which you adapt to other scenarios for your student project.

Step #1: Some relation exists between explanatory variables and a dependent variable.

- To fit a regression line, select explanatory variables or adapt them (e.g., take logarithms or square roots).
- Use the adjusted R^2, parameter significance, *p* values, and the principle of parsimony.

Step #2: Do the analysis separately for two similar samples. Join the samples and see if the same regression equation might be appropriate for both.

Alternatively, do the analysis for a large sample. Divide the sample into two homogeneous groups, and see if the same regression equation is appropriate for both groups.

Jacob: What do the illustrative worksheets show? Do we use these worksheets for our student project?

Rachel: The project template suggests many ideas using sports statistics. You are not bound by these suggestions, but they help you formulate ideas for your student project.

- The illustrative worksheets compare regression equations for two teams.
- The analysis is kept simple, with 100 data points, two teams, and short equations.
- The worksheets are carefully documented, so you can reproduce the analysis.

We provide won-loss records for other teams, sports, and years on the discussion forum. Sports statistics are on many web sites; use data from the discussion forum or other sites.

Jacob: The illustrative worksheets use Boston and New York. Does a student project use just two teams or all teams?

Rachel: You choose the teams (all teams, one League, or one team), years, and sport.

- You may test whether one team, such as the Boston Red Sox or the New York Yankees, differs from other teams. Use an *F* test to compare this team with the others.
- You may test if one team (say New York) differs between 1901-1960 and 1961-2007.

Take heed: If possible, use large enough samples that your results are significant.

Illustration: Your student project might use all teams with a dummy variable for the League to test if the two Leagues have the same regression equation.

Take heed: Your student project may use other statistics, such as batting averages, home runs, pitching statistics, or similar items in other sports (assists, points, rebounds).

Jacob: Do we use a dummy variable with D = 1 for one team and D = 0 for other team(s)? Or do we use a separate regression analysis on each team?

Rachel: One method uses dummy variables for the *F* test. Another method uses separate regression equations.

Take heed: Before doing the statistical test, examine the regression equation for one team vs the regression equation for other teams (or any other dichotomy).

~ If the regression equations differ much, we should need separate equations.
~ If the regression equations are similar, one equation is probably sufficient.

SETTING UP THE WORKSHEET

{If you are experienced with Excel, you don't need detailed instructions. This section is for candidates who rarely use Excel.}

Jacob: How do we set up the regression worksheet?

Rachel: The procedure for your student project depends on the hypotheses you test and the analysis you perform. Before setting up the data:

~ Replicate the analysis for Boston and New York. Be sure you can run the regression analyses with the *REGRESSION* add-in.
~ Decide what hypothesis (what *F* test) you will examine, which sport, which teams, and which years. You may modify your choice (or change it entirely) as you proceed.
~ Choose data for one team and set up an array. Do the regression analysis on this array first, to verify that you understand the procedure.
~ Modify your array for any errors that you find.
~ Set up arrays using all the data you need and complete the student project.

Step #1: *SELECT DATA*

Copy the data you need from the NEAS web site (sport, teams, years). Sort the data base by team, and form a regression for one team.

- Use years in *descending* order to replicate the format on the illustrative worksheet.
- The work-sheet for New York has simpler cell formulas. The worksheet for Boston uses blank cells instead of zeros. Use whichever form you prefer.

Take heed: Excel has excellent sort facilities, but it is easy to make errors if you are not careful. Save your spread-sheet before sorting. If you make an error, go back to the saved spread-sheet. A common error is to select the column of team names and sort them. That sorts just the names, matching them with the wrong data.

- Select the entire data base before sorting.
- Save the pre-sorted data in another file or worksheet.
- After sorting, compare two or three rows to ensure that your sort worked.

Illustration: For the baseball data base for 1901 – 1960, copy the losing percentages for Boston or New York. You copy a column of 60 figures.

Step #2: *CREATE A WORKSHEET*

In a new worksheet, place the cursor in the upper left cell of the new array. Leave room at the top of the spread-sheet and on the left-hand side for labels and comments.

Take heed: Document your student project. Course instructors have difficulty reviewing student projects that are not documented, and they may return your submission for clearer explanation. To ensure quick turn-around for your student project, document your work.

The course instructor reads the documentation in your write-up. If you begin the write-up after finishing the analysis, explaining each step of the student project is tiresome. Instead, document your analysis with comments or call-outs or text entries in nearby cells. When you complete the analysis, copy the documentation into a Word file (or other text file), edit the comments into a well-written report, and submit it to the NEAS office.

Illustration: The illustrative worksheet uses cell F11 as the upper left cell for the data. Five columns on the left and ten rows on top are left for labels and headings.

Step #3: *ROWS TO COLUMNS*

Create a matrix in the format shown on the illustrative worksheets. You can create the matrix various ways; use whatever is easiest for you.

- Your student project may use other data and a different format. The instructions here are for candidates who have not previously used the *REGRESSION* add-in.
- The data may be formatted by columns instead of by rows. The triangle for a single team is symmetric in rows vs columns.

An advantage of Excel over other spread-sheet software is the ease of reformatting the data. We mention how to convert columns to rows (or vice versa).

- The New York work-sheet forms the triangle by columns, without using *PASTE SPECIAL*.
- The Boston work-sheet forms the triangle by rows, using *PASTE SPECIAL*.

We can make the data matrix just as easily without converting columns to rows, but the Excel commands help you elsewhere as well.

Select *PASTE SPECIAL* from the *EDIT* menu. Select transpose from the screen of options. This transposes the column of losing percentages to a row. You should have a row of 60 figures from Cell F11 to Cell BM11.

Column F has the forecasts; Columns G and subsequent have the explanatory variables. The statistical techniques help select the optimal number of past years.

Take heed: The matrix is *symmetric* in rows vs columns. We repeat the instructions below switching rows and columns. The illustrative worksheets use *PASTE TRANSPOSE* for Boston, not for New York.

Step #4: *CREATE 50 TO 60 ROWS*

The regression worksheet uses 50 to 60 rows. Start with 60 rows, and either

- Convert cell formulas to values, and delete (or hide) the rows you don't use.
- Specify the only a portion of the data matrix for the *REGRESSION* add-in.

Take heed: By transposing columns to rows, we can delete rows without problems.

- For the New York worksheet, we re-typed three values after deleting rows.
- For the Boston worksheet, we transposed columns to rows.

We want to copy the row of losing percentages 59 times and shift the figures one column to the left in each new row.

- The first row has the 1960 losing percentage in Column F, 1959 in Column G, etc.
- The second row has 1959 in Column F, 1958 in Column G, etc.
- The third row has 1958 in Column F, 1957 in Column G, etc.

To do this efficiently:

- Type "=G11" in Cell F12.
- Copy Cell F12 to Cells G12:BM12.
- Copy Cells F12:BM12 to Cells F13:BM60.

This procedure puts zeros in cells that don't have values. If you want blank cells instead of zeros but you want to keep numeric formats in these columns, use an IF statement.

- Type =IF(G11="","",G11) in Cell F12. Alternatively, type =IF(G11<>"",G11,"")
- Copy Cell F12 to Cells G12:BM12.
- Copy Cells F12:BM12 to Cells F13:BM70.

The formulas give a matrix of 60 rows by 60 columns: losing percentages for 1960 – 1901.

We don't forecast the first decade of won-loss records, since newly formed teams may differ from established teams.

- If your data sample does not have many points, use all available years. The *F* test for New York vs Boston uses one past year, so we could use 1960 – 1902 (59 years).
- Your student project may compare regressions for new teams vs established teams.

Illustration: If you use years 1981-2005 for teams formed in 1981, use all available years.

In column E, enter the row labels. The first row forecasts year 1960, so place 1960 as the row label. The second row forecasts year 1959, so place 1959 as the row label.

Select cells E11 and E12. Drag the lower right corner (the *fill handle*) of cell E12 down to cell E70. Excel fills in the cells with 1958 through 1901.

You don't have to use Excel's autofill function. You can insert column headings and row labels manually. But *be sure to label your tables*. Your write-up should specify the years and teams in your project. Our faculty can not review an undocumented student project.

Use columns A, B, C, and D for additional row labels. You might use

- Column D for the team.
- Column C for the League, division, conference, or other group.
- Column B for the observation number.

After forming the data file for your student project with all the teams for your sport, sort the rows into any order you like.

Illustration: Your student project compares the regression equation for 1901-1960 with the equation for 1961-2005 for 8 teams with data in all 105 years.

- Form data triangles separately for each team and join them, using the procedures in the illustrative worksheets for Boston, New York, and the combined regression.
- Sort the worksheet by year. The sorted worksheet has 8 rows for 2005, 8 rows for 2004, and so forth.
- Use the rows for years 1961-2005 for one regression and 1901-1960 for the other.

Take heed: For some hypotheses in the discussion forum postings, the data points in each group are determined from the won-loss records.

Illustration: Your student project compares the regression equations for better teams vs worse teams. The quality of a team changes each year, depending of its losing percentage.

- Better teams: Losing percentage in previous year is less than 50%.
- Worse teams: Losing percentage in previous year is more than 50%.

Use the following steps:

- Form data triangles separately for each team and join them (as above).
- Include a dummy variable D = 1 if the losing percentage in the previous year < 50% and D = 0 if the losing percentage in the previous year > 50%. Decide whether a losing percentage = 50% is good or bad (or eliminate those rows).
- Sort the worksheet by the dummy variable. Continue as shown on the illustrative worksheet for the *F* test.

Note: The sorting is not necessary. If you use dummy variables (vs separate regression equations), you can do the entire student project with restricted and restricted equations.

But sorting the rows into better and worse teams and examining the correlations and regression equations for each group makes the student project clearer.

Jacob: What are the column captions in row 10?

Rachel: Enter *Year 0* or *Forecast* in cell F10; this is the forecast year. Enter *Year -1* in cell G10; this is the first prior year. Enter Year -2 in cell H10; this is the second prior year.

Select cells F10:G10 and drag the autofill handle on the bottom right of Cell G10 through cell P10. The column headings are the variable names in the *REGRESSION* output. In your write-up, explain which columns are used in each regression.

NEW YORK (AMERICAN LEAGUE) ILLUSTRATIVE WORKSHEET

The illustrative worksheet for New York copies a column of losing percentages from the initial data source. It then copies this column to 59 more columns, moving the losing percentages up one cell in each adjacent column.

Step #1: Copy the won-loss records for New York from the worksheet “WLRs.” Copy a range of 4 columns × 60 rows. The columns are

- Team = New York
- League = American League
- Year = 1960 to 1901
- Won-loss record = 37.0% to 38.9%.

Step #2: Create a new worksheet and name it New York (or whatever you use). Select Cell C11 and paste the value with Cntl-V. The range occupies the Cells C11:F70. The losing percentages are in Column F, rows 11 to 70 (F11:F70).

Take heed: You may have to switch Columns D and E

Step #3: Fill the rest of the matrix by formulas.

- In Cell G11, type "=F12"
- Copy Cell G11 to Cells H11:BM11 *Take heed:* You may see zero’s or blanks in these cells, depending on the cell formatting in your worksheet.
- Copy Cells G11:BM11 to Cells G12:GM70 *Take heed:* Figures should appear in the upper left triangle of your matrix. If the values are still zero, press F9 to re-calculate.

You created the full table needed for the regression analysis. The illustrative worksheet doesn’t use Rows 61 through 70 (the first 10 years), since the relations between years may differ for new teams vs established teams.

Take heed: The actual results do not differ much if we use all years. You can even do a student project comparing recently formed teams and established teams. Use won-loss records from the most recent 30 years, when teams have been added in all four sports.

Take heed: Your data table shows 0% in many cells if the columns are formatted as percentages. If you want blank cells instead of 0%, use the formula =(IF F12=“”, “”, F12). Enter this formula in cell G11 and copy it to H11:BM11.

Step #6: *EXPLANATORY VARIABLES AND THE REGRESSION ADD-IN*

Jacob: The rows have different lengths. Do the explanatory variables differ in each row?

Rachel: The explanatory variables are the same for each row. Use the *REGRESSION* add-in to select the explanatory variables. Suppose we forecast years 1911-1960 (50 years).

The Y variable is cells F11:F60. For the regression add-in, type F10:F60 in the text box for the Y variables and select *HAS LABELS*. This uses the entries in row 10 (cell F10) as variable labels, not as data.

Take heed: You can use F11:F60 and de-select *HAS LABELS*. The *REGRESSION* add-in gives generic names to the variables. Overwrite these names with other names.

Jacob: Can we select the cells instead of typing F10:F60?

Rachel: Yes, you can select the cells in the Excel worksheet.

Jacob: The *REGRESSION* add-in dialogue obscures the worksheet; how can I see the cells?

Rachel: Click the minimize button at the right of the entry box. The *REGRESSION* add-in dialogue collapses to a single line of that entry box only. Drag the reduced dialogue box to the top or bottom of the window so you can select the cells you want. Many Excel built-in functions and dialogues have this minimize button at the right side of each entry box.

Jacob: What are the X variables for the *REGRESSION* add-in?

Rachel: The X variables depend on the number of past years.

Start with one past year. The X variables are in cells G11:G60. Type G10:G60 and select *HAS LABELS*. You don't need residual output or graphs for this part of the student project.

- It is easiest to place the output on the same work-sheet, such as cell A75.
- Keep the regression output for each set of past years. Don't over-write the output in the same location.

Take heed: By default, Excel places the output on a separate worksheet to avoid over-writing cells. If you use separate worksheets, document their relation in your write-up.

Next use two past years. The X data are in cells G11:H60. Compare the regression statistics for two past years with those for one past year.

Take heed: If you use 2 or more explanatory variables, they must be in adjoining columns.

Jacob: What regression statistics do we examine?

Rachel: Examine the adjusted (or corrected) R^2. If another year does not materially raise the adjusted R^2, don't add the year.

Jacob: Why not state this in the affirmative, as *if another past year materially raises the adjusted R^2, add the year*?

Rachel: We have two more tests.

- If another past year causes some regression coefficients to be negative, we hesitate to add the year, unless we have an intuitive rationale for a negative coefficient.
- If the *t* statistic for the oldest year is low, we hesitate to add the year.

Jacob: The term *hesitate* doesn't sound objective.

Rachel: These are subjective decisions.

~ Raising the adjusted R^2 from 45% to 45.1% is not material. A change from 45% to 46% is less clear.
~ A *t* statistic of 0.25 is not significant. A *t* statistic of 0.75 is less clear. A *t* statistic of 1.75 is strong enough to include the extra year.

Jacob: At a 5% significance level, the *t* statistic should be 2 or higher.

Rachel: We use less stringent tests. Even if a β parameter has a *p* value of 15% to 25%, we *might* include the past year in the regression equation.

Jacob: Should we do this separately for each team, or for all teams together?

Rachel: Using all teams, or all teams in one League (or division), gives more data points and more significant results. But the choice of one team vs a group of teams depends on the student project.

- Baseball has many years, and we get reasonably efficient results even for one team.
- For sports with fewer years and higher stochasticity, such as hockey and football, a single team's experience may not give significant results.

Illustration: A student project on baseball might use a single team and compare two sets of years. A student project on football might compare two conferences.

Step #7: Regression Analysis

Derive the optimal number of years for the regression equation.

- This section reviews the Excel worksheet for Boston, American League, 1901-1960.
- See also the call-outs on the worksheet for New York, American League, 1901-1960.

Take heed: Your project may use other teams or years, but the concepts are the same.

Boston is an average team, with a 60 year losing percentage (won-loss record) of 50%. (The exact won-loss record is 49.9%.) We show the *REGRESSION* output for 1, 2, 3, and 4 past years. As you read this, check the four blocks of regression output.

We use forecasts for 1911-1960, since the first several years (before the teams became established) may have abnormal results.

Take heed: If your student project uses recent years, you may have both new teams and older teams in your data sample. You may compare the new teams vs the older teams to see if they have different regression equations.

Using only one past year gives an R^2 of 44.6% and an adjusted R^2 of 43.4%. The fitted regression equation is Y = 16.53% + 67.17% × X. The means of X and Y for these years are approximately (not exactly) 50%. Note that 16.53% + 67.17% × 50% = 50.12%. If we use all teams in a League, the means are exactly 50%.

Take heed: The regression line passes through the averages. To check the work, compute the average X and Y values, and verify that the regression passes through the averages.

The standard error of the independent variable is 10.8% and the *t* statistic is high (6.216). The *p* value is $1.17 \times e^{-7}$, or effectively zero. We summarize the regression results as:

- One third of the team's won-loss record is the overall average: 16.53% / 50% ≈ 33%.
- Two thirds (67.17%) reflects the quality of the team in the previous year.

Using two past years raises the R^2 to 51.8% and the adjusted R^2 to 49.7%. The standard errors of the independent variables increase to 13.7% and 13.6%. The ordinary least squares estimators are 42.9% and 36.0%, and their *t* statistics decline to 3.13 and 2.64.

Jacob: Why are the standard errors are about the same for each past year?

Rachel: The standard error of the ordinary least squares estimator is the standard error of the regression divided by the standard deviation of the losing percentages.

- Of the 50 losing percentages for the first prior and second prior years, 49 are the same.
- The standard deviation of the losing percentages are about the same for both years.

Illustration: We use two past years to forecast won-loss records for 1911-1960.

- X_1 is the losing percentages in 1910-1959.
- X_2 is the losing percentages in 1909-1958.

The standard deviations of these two explanatory variables are almost identical.

Jacob: Why do the standard errors of the ordinary least squares estimators increase between one past year and two past years? The standard deviation of the independent variable hasn't changed, and the standard error of the regression decreases from 7.47% to 7.05%. Shouldn't the standard error of the ordinary least squares estimators decrease?

Rachel: The two past years are highly correlated. The correlation with a lag of one year (ρ) is about 67%. The multicollinearity raises the standard error and lowers the *t* statistics.

Both past years are highly significant (*p*-values of 0.3% and 1.1%), so we prefer two past years to one past year.

Take heed: The project template explains the multicollinearity. You are not required to analyze the multicollinearity in your student project.

Jacob: Can you say in actuarial terms what these regression equations mean?

Rachel: If we don't know last year's won-loss record, the best estimate of this year's won-loss record is 50%.

- If we know last year's won-loss record, we assign it 67.17% credibility.
- The overall 50% won-loss record gets the complement, or 1 – 67.17% = 32.83%.

Jacob: The overall won-loss record is 50%, so (1 – 67.17%) × 50% = 16.42%. The regression equation has α = 16.53%. What causes this difference?

Rachel: The overall won-loss record for Boston is 49.9%, not 50%.

- Actuarial credibility assumes the true overall won-loss record is 50%, and the slight difference in the long-run average for Boston is random fluctuation.
- Regression analysis assumes the long-run average is the true expected overall won-loss record.

Take heed: Contrast the regression for New York, whose 60 year losing percentage is 42%, not 50%.

Jacob: Does a third past year improve the regression?

Rachel: Using three past years does not help the regression. The R^2 increases slightly to 52.6%, but the adjusted R^2 declines from 49.7% to 49.5%. This tells us that the increase in R^2 reflects the additional independent variable, not better explanatory power.

The standard error of the regression remains the same. The additional multicollinearity raises the standard errors of the ordinary least squares estimators and reduces their *t* statistics. The third past year is not significant, with a *p*-value of 36.8%. Two past years gives the optimal regression equation.

It is useful to check a fourth year, to verify the results. Using four past years gives a negative regression coefficient for the fourth year that is not significant at all, with a *p*-value of 45.8%. The adjusted R^2 declines further to 49.0%.

Take heed: The Boston illustration shows the various items to examine:

- The change in the adjusted R^2, not the simple R^2.
- Standard errors, *p* values, and signs of the β coefficients.

Jacob: Can you explain this? Why would another year hurt the regression?

Rachel: Consider a simpler scenario: the optimal regression is $Y = 25\% + 50\% \times X_1$.

- A team with a 60% won-loss record last year expects a 55% won-loss record this year.
- A team with a 40% won-loss record last year expects a 45% won-loss record this year.

Now suppose we also use a second past year. A won-loss record of S-T mean S% two years ago and T% last year. Consider two scenarios, W and Z. In Scenario W:

- A team with a 60%-60% past won-loss record expects a 57% won-loss record this year.
- A team with a 40%-60% past won-loss record expects a 53% won-loss record this year.

The second past year adds to our information about the team's quality. We might expect this in (American) football. Teams play only a dozen games each season, and random fluctuations (injuries, fumbles) greatly affect team performance. Last year's won-loss record may be higher or lower than expected because of random fluctuations that do not repeat. A second past year helps us predict the current year's performance.

Scenario Z is different:

- A team with a 60%-60% past won-loss record expects a 55% won-loss record this year.
- A team with a 40%-60% past won-loss record expects a 55% won-loss record this year.

Last year's won-loss record gives all the relevant information.

- A 60% won-loss record reflects the teams' quality. It differs from the won-loss record two years ago because team quality changed, not because of random fluctuations.

- The second past year adds noise to the regression analysis, not information. Regression analysis filters the data to extract the information and eliminate the noise.

Jacob: The instructions use ten past years. The illustrative worksheets use two past years for Boston and one for New York. Do we expect to use just one or two past years, or are these two teams exceptional?

Rachel: The answer depends on the sport: number of games in the season, number of players on the team, average playing lives, draft rules, and the stochasticity of the sport.

- You never need more than ten years.
- Most student projects are simpler, and 2, 3, 4, or 5 past years are sufficient.

Jacob: Do we expect two past years to be optimal or should we expect 5 past years?

Rachel: The results differ by sport. Baseball has a long season: two years is now 324 games. (It was fewer before 1961, but still many games.) This is about 20 football seasons.

DATA POINTS AND FORECASTS

{This dialogue notes that data points = number of forecasts × number of teams.}

Jacob: How many forecasts does each team have?

Rachel: If each forecast needs 10 past years, the first forecast is for 1911 and the last forecast is for 1960, so each team has 50 forecasts.

Jacob: With two Leagues and 8 teams per League, we have 16 × 50 = 800 data points. Do we need so many points?

Rachel: The student project demonstrates that you can apply statistical techniques to real data. The data on the NEAS web site is in Excel arrays formatted for the project template, so you can easily copy enough data points for accurate estimators.

You do not use all the data on the web site. We included all the teams for four sports because candidates might prefer a student project on their home teams.

Use the following guidelines for the number of data points you should use:

~ More data points give more accurate results. If you use data from the NEAS web site, use 80 to 400 points.
~ If you use a subset of teams, you may normalize the data. If you compare Teams A and B, and their overall won-loss records are 40% vs 60%, you should normalize both teams to a 50% overall won-loss record. The illustrative worksheet for Boston vs New York show how to normalize the won-loss records.
~ *Note:* Normalization is not covered in the course. We do not grade your student project adversely if the data are not normalized. Look at the illustrative worksheet and decide if you should normalize your data.
~ If you compile your own data, it may be hard to get 80 points. For sports with few data points (few years or few games per year), such as women's soccer, pay close attention to the standard errors of your estimators. Try to have at least 20 or 30 points, but don't be concerned about possible distortions from random fluctuations. We grade the student project on the quality of the analysis, not on the number of data points.

OPTIMAL NUMBER OF YEARS

{This dialogue discusses the optimal number of past years as explanatory variables. The illustrative worksheet shows the regression equations for one team and 50 data points. You use a similar worksheet for your student project.}

Jacob: If the correlations are positive for the first 9 lags, do we perform the regression analysis using 9 past years (independent variables) and one constant term?

Rachel: Do the regression analysis step by step.

- Use first 1 past year and determine the adjusted R^2.
- Then use 2 past years and determine the adjusted R^2.
- If the adjusted R^2 increases, use 3 past years and determine the adjusted R^2.

Continue in this fashion until the adjusted R^2 stops increasing.

Jacob: If the first nine correlations are positive, won't the adjusted R^2 increase for 9 years?

Rachel: No. Suppose the true relation is $Y = \alpha + \beta \times X + \epsilon$ with $R^2 = 81\%$.

- The correlation of lag 1 is 90%.
- The correlation of lag N is $90\%^N$.

The optimal regression equation uses one independent variable, but the correlations are positive for many years. Even for the tenth prior year, the correlation is 34.87%. The adjusted R^2 is highest for one independent variable.

Take heed: If you have few data points and high stochasticity, the adjusted R^2 may be higher for two or three past years because of random fluctuations. With several hundred data points, random fluctuations are smoothed.

Jacob: Do we use the regression equation with the maximum adjusted R^2?

Rachel: If the adjusted R^2 is 40%, 55%, 60%, and 59% for 1, 2, 3, and 4 past years, use 3 past years. But the pattern is not always clear.

Illustration: If the pattern is 40%, 55%, 60%, 63%, 65%, 66%, 66.5%, we might stop at 65%, or 5 past years. Pay close attention to the regression coefficients. If using 6 past years gives some negative regression coefficients, such as –3% for year N-5 and +2% for year N-6, don't use the regression equation with 6 past years.

When in doubt, use the principle of parsimony. If you are unsure whether to use N past years or N+1 years, use N years. Explain in your write-up that the two alternatives seemed equally good and you chose fewer years.

Jacob: Why might we get negative regression coefficients? If each past year has a positive correlation with the forecast year, aren't the regression coefficients all positive?

Rachel: If the independent variables are *orthogonal*, the regression coefficients are all positive. If the adjusted R^2 is 40% for one past year, the correlation of adjoining years is $\sqrt{40\%} \approx 63.25\%$. With high multicollinearity, the β estimators have a high standard error.

Illustration: If the true β is low, such as 2%, and the standard error is 3%, the ordinary least squares estimator is sometimes negative.

Take heed: Multicollinearity and stochasticity cause negative regression coefficients.

Illustration: Suppose the regression coefficients are 50% for lag 1 and zero for other lags. The standard deviation of the ordinary least squares estimators is high because the won-loss records of the past years are correlated with each other. If the standard deviation is 10%, we might observe values of –15% or –20% from random fluctuations.

Note: The time series on-line course discusses this topic. You use Bartlett's test to judge the significance of the regression coefficients. For the regression analysis student project, assume a negative coefficient reflects random fluctuation.

Jacob: How do we decide how many years to use?

Rachel: Much of this is subjective. Use the following procedures:

~ If the *correlation* of lag *k* is close to zero, ignore years *k* and older.
~ We expect positive correlations for more years than we use in the regression equation.
~ If the *adjusted* R^2 does not increase, the added year does not help. A decrease in the adjusted R^2 is unambiguous; a slight increase is unclear. If the increase is not material, you don't need the additional year.
~ If a regression coefficient is *negative*, we are probably using too many years. Avoid regression equations with negative coefficients.
~ If the *t statistic* for a regression coefficient is not significant, ignore the year. Use a loose significance level, such as a *p* value of 10% or 20%.

The phrases *close to zero*, *increase*, *not significant* are subjective. You decide how strong a significance level to use.

Illustration: Suppose the coefficients for the first four lags are positive, but the *p* value for the fourth past year is 35%. It is not clear if we should use 3 years or 4 years. If your write-up, explain how you chose the number of years.