SPORTS SCORES PROJECT TEMPLATE F TEST

(The attached PDF file has better formatting.)

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 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 vs those west of the Mississippi, 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. For the *F* test, use the same number of years in each group.

Illustration: The optimal regression equation uses five years for one group of teams and six years for the other group of teams. You can use regression equations with either six or five past years for the *F* test. The student project checks if you understand how to apply an *F* test, not if you have found the optimal regression equations.

Take heed: 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.

Take heed: If we use rich teams vs poor teams or good teams vs bad teams, the average won-loss records differ. If you do not normalize the won-loss records, the null hypothesis fails the *F* test. 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 result is a probabilistic statement: "If the null hypothesis is true, the probability of obtaining the observed (or more extreme) results is Z%…" The conclusion may be: "Using a 10% significance level, I infer that the same regression equation may be used for both Leagues."

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:

Better vs Worse Teams: 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.

Sport: 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 11^{th} draft pick and the 12^{th} draft pick does not have a material effect. This implies that good teams and bad teams have different regression coefficients.

F TEST BACKGROUND

Review the textbook chapter and course module #11 on the *F* test. Suppose your student project compares the regression equations for the American vs the National League.

- For the American League, the regression equation is $Y = \alpha_1 + \alpha_2 X_2 + \alpha_3 X_3 + \alpha_4 X_4 + \epsilon$.
- For the National League, the regression equation is $Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_3 X_3 + \epsilon$

Use the same number of past years for both Leagues.

Let *D* be a dummy variable for which American League = 0 and National League = 1.

The constrained equation is $Y = \alpha_1 + \alpha_2 X_2 + \alpha_3 X_3 + \alpha_4 X_4 + \epsilon$.

The unconstrained equation is

$$Y = \alpha_1 + \alpha_2 X_2 + \alpha_3 X_3 + \alpha_4 X_4 + (\beta_1 - \alpha_1) D + (\beta_2 - \alpha_2) D X_2 + (\beta_3 - \alpha_3) D X_3 + (\beta_3 - \alpha_3) D X_3 + \epsilon$$

This equation seems to have four independent variables (D, X_1, X_2, X_3) and a constant (α_1). But a regression equation must be linear *in the coefficients*. It actually has 7 independent variables ($D, X_1, X_2, X_3, D \times X_1, D \times X_2, D \times X_3$) and a constant.

Take heed: The *F* test has several (mathematically equivalent) forms. Review course module #11. This posting uses a dummy variable and the error sum of squares (ESS). You can also use the forms with R^2 or the regression sum of squares (RSS).

Suppose we have eight teams in each league, we forecast ten years, and the error sums of squares are

- American League: $ESS_A = 4$
- National League: $ESS_N = 8$
- Constrained equation: $ESS_C = 15$

Note that $4 + 8 = 12 < 15$: that is, the regression equation fitted for the American League is not identical to the regression equation fitted for the National League.

The null hypothesis is that the two equations are the same. The apparent difference stems from random fluctuation, not from an inherent difference between the leagues. The *F* statistic is the probability that the improvement in the ESS reflects the greater degrees of freedom and random fluctuation.

DEGREE OF FREEDOM AND RESTRICTIONS

The unrestricted regression has 80 data points and 8 explanatory variables for $80 - 8 = 72$ degrees of freedom. It has four restrictions:

- $(\beta_1 - \alpha_1) = 0$
- $(\beta_2 - \alpha_2) = 0$
- $(\beta_3 - \alpha_3) = 0$
- $(\beta_4 - \alpha_4) = 0$

Take heed: The restrictions are the same whether or not we use dummy variables.

The four restrictions are

- The constant term is the same for the American League as the National League.
- The β for the first prior year is the same for the two leagues.
- The β for the second prior year is the same for the two leagues.
- The β for the third prior year is the same for the two leagues.

We compute the *F* statistic:

- The difference in the ESS between the restricted and unrestricted regressions is 15 – 12 = 3.
- The ESS for the unrestricted regression is 4 + 8 = 12. The ESS divided by the degrees of freedom is 12 / 72 = 1/6.
- The value of the *F* statistic is { 3 / 4 } / { 12 / 72 } = 4.5.

The *F* statistic follows a distribution that depends on its degrees of freedom. The *F* statistic has two degrees of freedom, one for the numerator and one for the denominator.

If the *F* statistic is greater than the critical value for a given level of significance, we reject the null hypothesis and presume that *at least one* of the restrictions is not correct. That is, at least one of three past years or the constant term has a different effect for the American League vs the National League.

SIGNIFICANCE LEVEL

Choose and justify a significance level, reflecting your assumptions and objectives.

- If you assume the two sets of teams differ materially, use a loose significance level.
- If you assume the two sets of teams do not differ, use a stringent significance level.

Illustration: We compare the American League with the National League. We assume the two leagues are not truly different, and observed differences reflect random fluctuations in the won-loss records. We reject the null hypothesis only if the empirical evidence is strong that the two leagues differ: that is, at a stringent significance level.

Illustration: We compare the basketball team that received the first draft pick with the team that received the last draft pick. (These teams change each year.) We presume the won-loss record changes greatly when a team receives the first draft pick, so the β is higher for the team that received the last draft pick. We reject the null hypothesis even at a less stringent significance level.

Intuition: Suppose we have only 20 data points and a moderate difference in the regression equations, giving a *p* value of 15% for the *F* test.

- We don’t reject the null hypothesis that the two leagues are the same. We need stronger empirical evidence to reject our null hypothesis.
- We reject the null hypothesis that the draft pick number has no effect. We assumed this hypothesis was not true, and a *p* value of 15% is enough to confirm our belief.

Your student project write-up should justify your significance level and your conclusions.

F STATISTIC

{The *F* test can be applied to various hypotheses, which differ for each student project.

The write-up should state the hypothesis that you are testing, what the *F* test measures, and what you conclude. This dialogue explains the *F* statistic, degrees of freedom, and hypothesis testing. It shows the formulas and references the pages in the textbook.}

Jacob: What does the *F* statistic show?

Rachel: The *F* statistic shows if a group of regression coefficients are significant.

Jacob: Do you mean to say: "If all the β parameters as a group are significant"?

Rachel: One use of the *F* statistic is to test whether all the β parameters as a group are significant. The student project is a more sophisticated use of the *F* statistic.

Take heed: The ANOVA (analysis of variance) portion of the *REGRESSION* add-in gives an *F* statistic and its significance. This is not the *F* test you use in the student project.

Illustration: Suppose we test if the optimal regression equation is the same for National vs American League. We write the regression equation so a group of parameters measures the difference between the Leagues. We test if this group of parameters is significant.

The *F* statistic can be used to test more general hypotheses. A more precise statement is that the *F* statistic tests any linear relation among the regression coefficients.

~ One linear relation is that $\beta_j = \beta_k = 0$.
~ Another linear relation is that $\beta_j + \beta_k = 0$ or $\beta_j - \beta_k = 0$.

Take heed: You can choose any topic for your student project, develop a hypothesis about the similarity or difference of two samples, and use an *F* test to verify the hypothesis.

Jacob: How do we calculate the *F* statistic?

Rachel: Read sections 5.3.1 (pages 128-131) and 5.3.3 (page 133-135). These are two perspectives on the *F* test; either perspective is fine for the student project.

- Get the figures for the *F* statistic from the *REGRESSION* add-in.
- Calculate the ratio with a cell formula.

Take heed: The *REGRESSION* add-in shows an *F* statistic. We do not use this figure.

If you are unclear about the degrees of freedom in the numerator and the denominator and the error sum of squares in the restricted equation and the unrestricted equation, review the examples in Mahler's *Guide to Regression*. This saves you time on the student project.

Mahler's *Guide* is posted on the discussion board in 15 PDF files. Its practice problems cover the technical aspects of the *F* test. Once you have read a few practice problems, you should not have difficulty with this student project.

Take heed: All the formulas with examples from the illustrative worksheet are in this project template. Use the textbook or Mahler's *Guide to Regression* to understand the *F* test. The formulas for the student project are shown below.

Suppose we compare National vs American League teams. We test whether the estimated regression coefficients are the same for the two Leagues.

Take heed: The project template explains an *F* test for National v American League teams. The illustrative worksheet shows data for Boston vs New York (American League). This project template then computes the *F* test and discusses intercepts, slope coefficients, and normalization. Your student project applies an *F* test to another pair of data samples.

Hypothesis Testing and the F Test

{Actuaries solve for numbers, such as premium = $3,000 or reserves = $10 million. This figure may be uncertain, but the figure is used by the insurer: the policy costs $3,000 and the balance sheet liability is $10 million. The *F* test does not solve for a figure. It does not say: "The true regression equation is …" It gives a probabilistic statement: "If the two teams have the same regression equation, the probability of observing the sample data is …"}

Jacob: Solving for the ordinary least squares estimators gives different β parameters for each League. The probability that the estimators are exactly the same is very low.

Rachel: The ordinary least squares estimators are not the true regression coefficients.

- We are *not proving* that the regression coefficients are the same.
- Rather, we test the *null hypothesis* that the parameters are the same.

If the probability of observing the sample data is less than a specified significance level, we reject the null hypothesis; otherwise, we do not reject the null hypothesis.

Take heed: This is true for all hypothesis testing. We never prove that the null hypothesis is true. We examine the *probability* of obtaining the observed data (or more extreme data) *if the null hypothesis is true*.

Jacob: Suppose that our analysis of the correlations and regressions suggests we should use eight past years to predict future won-loss records. Do we use eight past years for the *F* test as well?

Rachel: The Excel *REGRESSION* built-in function allows up to 16 independent variables. Eight past years gives 2 × (8 + 1) = 18 independent variables in the unrestricted regression equation. To keep your student project simple, use four or five past years. Even if eight years gives a better prediction, use four or five years for the *F* test in the student project.

Jacob: What if two or three years is optimal for each league separately? If we combine the two leagues, should we use more past years?

Rachel: If two or three years is optimal for each league (or division or conference or team or set of years) separately, use two or three years for the *F* test as well. The illustrative worksheet shows two years is optimal for the Boston baseball team and one year for New York. Use either one year or two years for an *F* test comparing these two teams.

Take heed: The Excel *REGRESSION* add-in needs a contiguous range for the explanatory variables. If you use N past years:

- Place the N past years in contiguous columns.
- Place the N+1 additional explanatory variables for the unrestricted equation in the next N+1 columns.

For clarity, the illustrative worksheet uses the *REGRESSION* add-in for one past year. For two past years, you must re-arrange the columns.

{*Note:* The *REGRESSION* add-in is written by a third party vendor, not by Microsoft. Its features differ from other Excel functions, and it has its own help files. We discuss certain aspects of the *REGRESSION* add-in to save you time on your student project.}

FORMULAS

Jacob: What is the formula for the F statistic?

Rachel: We show formulas with dummy variables and with separate regression equations. They are mathematically equivalent. Use any version for the student project.

Take heed: The illustrative worksheet for Boston and New York uses dummy variables.

Jacob: What is the *restricted* (or *constrained* or *reduced*) regression equation?

Rachel: The *restricted* regression equation assumes the null hypothesis is true, such as

- The regression equations for the National and American League teams are the same.
- The regression equations for Boston and New York are the same.

With two past years, the equation is

$$WLR_T = \alpha + \beta_1 WLR_{T-1} + \beta_2 WLR_{T-2} + \epsilon.$$

The restricted equation does not distinguish leagues, teams, divisions, etc. The same regression coefficients are used for all points.

Estimate the α and β coefficients with the *REGRESSION* add-in. The average won-loss record in any year is 50%, so $50\% = \alpha + \sum \beta_j \times 50\%$, or $2\alpha + \sum\beta_j = 1$.

Take heed: The 50% average won-loss record is one for a league (division, conference). A team or group of teams may have a different average. If your student project compares good teams with bad teams, you expect different averages.

Jacob: What is the *unrestricted* (or *unconstrained* or *full*) regression equation?

Rachel: We include a dummy variable: D = 0 for National League and D = 1 for American League. The unrestricted regression equation is

$$WLR_T = \alpha + \beta_1 WLR_{T-1} + \beta_2 WLR_{T-2} + D \times \alpha' + \beta_1' \times D \times WLR_{T-1} + \beta_2' \times D \times WLR_{T-2} + \epsilon.$$

Take heed: One dummy variable gives several additional regression coefficients. If the restricted equation has N independent variables and 1 intercept, the unrestricted equation has 2N+1 independent variables and 1 intercept.

Jacob: Won't the second equation always fit better?

Rachel: The question is whether it fits better because it uses more variables or it fits better because the National and American League teams differ.

Jacob: To answer that question, we compare the adjusted R^2 of the regression equations. What does the F test tell us that the adjusted R^2 doesn't?

Rachel: The adjusted R^2 is higher if the unrestricted equation is enough better to offset the greater degrees of freedom.

The adjusted R^2 gives two numbers, such as 44% and 46%, not a probabilistic statement. We can't test a hypothesis with the adjusted R^2.

Hypothesis testing assumes the null hypothesis has an a priori logic. We reject the null hypothesis only if the evidence is strong.

Illustration: The adjusted R^2 for the National vs American League may be

- 44% for the restricted equation
- 46% for the unrestricted equation

The *F* test may have a *p* value of 8%. If we see no reason for a difference in the leagues, we may not reject the null hypothesis. Random fluctuation may make the unrestricted equation fit better, even though the leagues do not differ.

Take heed: Your project write-up should relate the significance level to your assumptions. If the *F* test has a *p* value of 15%, your conclusion may depend on the scenario.

- When comparing baseball leagues: "I assume the National and American League are the same, so I do not reject the null hypothesis at a 15% *p* value."
- When comparing baseball vs (American) football: "These sports differ in games per season, players per team, and the number of seasons that players last. I do not expect the same regression equation, so I reject the null hypothesis at a 15% *p* values."

The *F* test gives the probability of observing the sample points if the restricted equation is true (e.g., the two sets of teams have the same regression equation).

Take heed: If the adjusted R^2 is lower for the unrestricted equation, the dummy variable does not help at all. We should not reject the null hypothesis for any significance level. The *F* test uses the simple R^2, which will be slightly higher for the unrestricted equation. The *F* value will generally be very low, and its *p* value will be close to one.

NUMERATOR AND DENOMINATOR

{The *F* test uses a *ratio of two ratios*, and it compares the result to a critical value. The dialogue below discusses each piece of the ratio and explain the intuition for the test.}

Jacob: What are the numerator and denominator of the *F* ratio?

Rachel: The denominator is the error sum of squares (ESS) for the unrestricted regression equation divided by the degrees of freedom. Alternative formulas use the RSS or the R^2.

- Learn first the *F* test with the ESS, which has the clearest intuition.
- The other formulas divide by the TSS and take complements.

Use the *REGRESSION* add-in to find ESS, RSS, and R^2.

Take heed: The residual SS in the *REGRESSION* add-in is the ESS.

Jacob: Which of these do we use: ESS, RSS, or R^2?

Rachel: The formulas are mathematically identical.

Take heed: Use the adjusted R^2 to select the optimal regression equation. Use the plain R^2 for the *F* test. The *F* test has a separate term for the degrees of freedom.

Jacob: How many degrees of freedom does the *F* test have?

Rachel: The *F* statistic has a degrees of freedom for the numerator and the denominator.

> The degrees of freedom for the denominator is the number of data points minus the number of explanatory variables in the unconstrained regression equation.

Illustration: T past years gives $2 \times T$ of β parameters and two α parameters. $T = 2$ gives 6 parameters. If we have N data points, we have $N - 6$ degrees of freedom.

Using years 1901-1960 and two past years in the regression, we can predict years 1903-1960 = 58 sets of years. With 16 teams, we have $16 \times 58 - 6 = 922$ degrees of freedom.

Take heed: Using dummy variables, the unrestricted equation has $2 \times T + 1$ independent variables and 1 intercept. The total number of variables is the same.

Take heed: You do not need to use all the data for the student project. Use enough data to get reasonable equations, but don't worry about missing some teams.

Take heed: The illustrative worksheet forecasts years 1911-1960, to avoid any differences for newly formed teams.

Take heed: With few data points, the degrees of freedom are important. With 928 data points, a slight error in the degrees of freedom won't change your conclusion. But your student project should demonstrate that you know how to compute the degrees of freedom.

Jacob: What is the numerator of the *F* ratio?

Rachel: The numerator is the difference in the error sum of squares (ESS) for the restricted and unrestricted equations, divided by the degrees of freedom for the numerator.

We use the *REGRESSION* add-in to find ESS, RSS, and R^2 for the restricted equation just as we did for the unrestricted regression equation.

Jacob: What are the degrees of freedom for the numerator?

Rachel: The degrees of freedom (q) are the difference in the number of parameters. These are α', β_1', and β_2', or 3 parameters. This is equation (5.20) on page 129 of the textbook.

$$F_{q,N-k} = \frac{(ESS_R - ESS_{UR}) / q}{ESS_{UR} / (N - k)}$$

More precisely, the degrees of freedom is the number of constraints. The null hypothesis has three constraints here:

- The intercept α is the same for the two leagues.
- β_1 is the same for the two leagues.
- β_2 is the same for the two leagues.

Using dummy variables, the constraints are

- The coefficient of D is zero.
- The coefficient of D × β_1 is zero.
- The coefficient of D × β_2 is zero.

For this analysis, the number of constraints is the difference in the number of parameters.

Jacob: How do we form the *F* ratio using R^2?

Rachel: Equation (5.21) on page 130 of the textbook shows the ratio using R^2. We can form a similar equation using RSS instead of R^2.

$$F_{q,N-k} = \frac{(R^2_{UR} - R^2_R) / q}{(1 - R^2_{UR}) / (N - k)}$$

Jacob: Why are these two formulas mathematically equivalent?

Rachel: The total sum of squares, TSS, is the same for the restricted and unrestricted regression equations. $R^2 = 1 - ESS/TSS$ (or $R^2 = \frac{RSS}{TSS}$). Substitute for R^2 in Equation 5.21 and factor out the TSS gives the Equation 5.20.

{If the formulas are confusing, review the practice problems in Mahler's *Guide to Regression*. Twenty minutes with the *Guide* may save you a hour with the student project.}

Take heed: Understand the rationale for the *F* statistic. Your write-up may state what the *F* test implies and how you chose the significance level. Don't write: "To see if the two leagues differ, I used an *F* test, which showed they have different regression equations."

- The *F* test makes probabilistic statements.
- Explain the probabilistic statement of the *F* test.

F TEST RATIONALE

{The *F* test seems abstruse, but its logic is simple.}

Jacob: What is the rationale for the *F* statistic?

Rachel: The error sum of squares depends on two things: the number of explanatory variables and the quality of the fit.

- If Equation Y fits better than Equation Z, its error sum of squares ESS is smaller and its regression sum of squares RSS is higher (by the same amount).
- If Equation Y has more explanatory variables than Equation Z but its fit is no better, its ESS is smaller and its RSS is higher (by the same amount).

Jacob: What does the second sentence above mean? If the error sum of squares is smaller, the equation fits better. What do you mean by "the fit of Equation Y is no better than that of Equation Z, but its error sum of squares is lower"?

Rachel: Suppose Equation Z has two explanatory variables. Equation Y adds a third explanatory variable, which is *unrelated* to the dependent variable.

- Equation Y does not fit better. When we use Equation Y or Equation Z to *forecast other values of the dependent variable*, the mean squared error is the same.
- Equation Y has a lower error sum of squares and a higher regression sum of squares. It has a higher R^2, though probably not a higher adjusted R^2.

The degrees of freedom for the *F* test teases apart these two items.

- The restricted equation has fewer explanatory variables, so it has a larger ESS.
- The unrestricted equation has more explanatory variables, so it has a smaller ESS.

The numerator of the *F* statistic says: "What is the average reduction in the error sum of squares for each additional explanatory variable?"

Jacob: That makes sense; what does the denominator do?

Rachel: Suppose we add two explanatory variables, which reduce the ESS from 500 to 400. We want to know if this is significant.

Intuition: We compare the percentage reduction in the ESS to the percentage reduction in the degrees of freedom. The change in the ESS above is a 20% reduction.

- If the degrees of freedom declines from 50 to 48, for a 4% decline, the ratio 20% / 4% = 5 is high, and the change in the ESS is material.

- If the degrees of freedom declines from 10 to 8, for a 20% decline, the ratio 20% / 20% = 1 is low, and the change in the ESS is not material.

For the intuition of the *F* test, think of the percentage ESS reduction as a ratio to the percentage reduction in the degrees of freedom. The formula for the *F* test is not written this way in the textbook; you must re-arrange the ratios.

The equations below show the *F* test using R^2. Do the same for the *F* test using the ESS.

$$F_{q,N-k} = \frac{(R^2_{UR} - R^2_R)/q}{(1 - R^2_{UR})/(N - k)} \qquad F_{q,N-k} = \frac{(R^2_{UR} - R^2_R)/(1 - R^2_{UR})}{q/(N - k)}$$

The intuition using the formula as written in the textbook is the same, but we need more words to express it.

- The numerator is the reduction in the ESS *per degree of freedom*: (500 – 400) / 2 = 50.
- The denominator is the ESS (in the unrestricted equation) *per degree of freedom*.

The ESS in the unrestricted equation is 500.

~ If we have 105 data points and 5 explanatory variables in the restricted equation, adding two more explanatory variables affect the ESS only if they actually explain some of the variance of the dependent variable. The *F* statistic is high.
~ If we have 10 data points and 5 explanatory variables in the restricted equation, adding two more explanatory variables has a spurious effect on the ESS. To see this, note that adding 5 more explanatory variables reduces the ESS to zero, since a total of 10 explanatory variables perfectly explains the 10 data points even if these variables are arbitrary. The *F* statistic is low.

Take heed: The two ways of expressing the intuition are equivalent.

Jacob: Must we explain the rationale of the *F* test in the write-up for the student project?

Rachel: No. Your write-up should have the following.

- Write the restricted and unrestricted equations.
- Explain the dummy variable (if you use one) or parameters with primes (α' and β_j').
- Explain the two forms of the null hypothesis:

  ~ An *intuitive* statement, such as "the regression equation is the same for the National League and the American League."
  ~ A *statistical* form, such as "the parameters α', $ß_1'$, and $ß_2'$ are all zero."

- Derive the degrees of freedom for the numerator and the denominator from the number of data points, the number of explanatory variables, and the number of constraints.
- State the form of the *F* statistic that you use (ESS, RSS, or R^2).
- State the significance level and critical values. You may write: "For a 10% significance level, the critical value with (N,D) degrees of freedom is …."
- Compute the numerator, denominator, and *F* ratio. Compare the F ratio to the critical value for a given degrees of freedom and significance level.

Take heed: The critical values are on pages 606-609 in the textbook for selected degrees of freedom. Many years ago, statisticians interpolated from hardcopy tables. Now Excel gives exact critical values for any degrees of freedom.

Illustration: Suppose we have 100 data points, 4 explanatory variables in the unrestricted equation, and 2 constraints. The degrees of freedom are (2, 96): 2 for the numerator and 96 for the denominator.

The critical F ratio for a 10% significance level is FINV(0.10, 2, 96) = 2.3587.

Alternatively, state the *p* value for your comparison. You may write:

For an F ratio of 2.16931 with (2, 96) degrees of freedom, the *p* value is FDIST(2.16931, 2, 96) = 11.98%, meaning that …" Explain if you reject or do not reject the null hypothesis.

Take heed: Critical values and hardcopy tables are from pre-computer years. Using the *p* value from the FDIST built-in function is the proper statistical practice.

Jacob: What do we expect to find from the F test?

Rachel: If there is no material difference in the two samples, such as Boston vs New York or National League vs American League, we do not expect to reject the null hypothesis, so the *F* statistic should be lower than the critical value. If the samples differ greatly, we expect to reject the null hypothesis, so the *F* statistic should be lower than the critical value.

Better stated: If the two samples do not differ materially, we expect a high *p* value, meaning that the observed differences reflect random fluctuation. If the samples differ materially, we expect a low *p* value: observed differences can not be attributed to random fluctuation.

Jacob: What would be an example where the two samples differ?

Rachel: Suppose the draft rules differed in the two Leagues: in one League, the worst team gets the first draft pick and in the other League, the best team gets the first draft pick. The two Leagues should have different regression coefficients.

Jacob: Can you give a more realistic example?

Rachel: Over the years, the draft rules and free agency rules have changed. We might compare two sets of years: before and after a change in the free agency rules. We use an *F* ratio to test if the regression equation is the same before and after the rule change.

Take heed: The change from the early days of each major league sport to current times is often discussed by sports commentators, though the truth of the statements is suspect. The common opinion is that in the early days of the sport, good teams stayed good for many years and bad teams stayed bad. But unequal teams reduces public enthusiasm for the game, so the draft (and other) rules have been changed to make teams more equal.

We don't know if this opinion is correct. To test it, see if the fitted regression equation changes from old years to recent years, with a lower β_1 in recent years.

Jacob: How do we calculate the *F* ratio to compare sets of years? Suppose we compare years before and after a change in the free agency rule.

Rachel: The procedure is the same for teams, leagues, years, players, or any criterion. For this illustration, we use two separate regression equations. Suppose we have

- (*k*-1) past years ⇛ *k* explanatory variables (including the intercept).
- *N* data points before the change in the free agency rule
- *M* data points after the change.

The number of data points is the number of forecast years times the number of teams. For this analysis, we might use two or three teams.

- The error sum of squares of the unrestricted regression equation is the sum of the ESS for the two separate regression equations.
- The degrees of freedom for the numerator are *k*
- The degrees of freedom for the denominator are the number of data points *N*+*M* minus the number of explanatory variables in the two regression equations, 2*k*.

Note: If T is the number of past years, the degrees of freedom in the numerator are T+1 and the degrees of freedom in the denominator are N + M – 2(T+1).

Jacob: How do we write this using dummy variables?

Rachel: Let the dummy variable D = 1 after the change in free agency and D = 0 before.

Jacob: In general, M ≠ N. In the previous examples (American League vs National League or Boston vs New York), we compare two data sets of the same size. Now we compare two data sets of different sizes. Does this cause a problem?

Rachel: The formulas are the same. We might compare New York vs other teams, which also compares data sets of different sizes.

Jacob: Is this formula in the textbook?

Rachel: This is Equation (5.25) on page 134:

$$\frac{(ESS_R - ESS_{UR}) / k}{ESS_{UR} / (N + M - 2k)}$$

Take heed: In the equation above, *k* is the number of explanatory variables including the intercept. For most statistical analyses, a difference of 1 in the value of *k* is not material. The opposite is true for the *F* test. For 1 past year, *k* is 2, not 1. Using 1 instead of 2 doubles the F statistic.

Jacob: What significance level should we use for the F test?

Rachel: The significance level is subjective; it depends on the number of data points and the intuition for the alternative hypothesis.

~ If we have many data points and no rationale for a difference in the two sets of data, such as National vs American League teams, we use a strong significance level, such as 2.5% or 1%. We reject the null hypothesis only if we are pretty certain it is false.
~ If we have few data points and a good rationale for a difference in the two sets of data, such as years before and after a large change in the free agency rule or the draft rules, we have a weaker significance level. We reject the null hypothesis even if we have only weak evidence that it is false.

F DISTRIBUTION VS X*-SQUARED DISTRIBUTION*

{This dialogue explains the relation of the *F* distribution and the χ-squared distribution.}

Jacob: We use an *F* test to compare two Leagues or two teams or two sets of years. In another statistics course, we used a χ-squared distribution. Are the *F* distribution and the χ-squared distribution the same? When do we use one vs the other?

Rachel: The F distribution and the χ-squared distribution are like the student's *t* distribution and the standard normal distribution.

~ We use the *z* statistic if we know the variance of the error term.
~ We use the *t* statistic if we do not know the variance of the error term and we must estimate it from the regression analysis.

The same applies to the F distribution and the χ-squared distribution.

~ We use the *F* distribution if we know the variance of the error term.
~ We use the χ-squared distribution if we do not know the variance of the error term and we must estimate it from the regression analysis.

F TEST IMPLEMENTATION

{This dialogue explains how to implement the *F* Test in your student project. The illustrative worksheets provide the cell formulas and VBA macros to use the *F* test. We grade the student project on your use of the statistical techniques, not your versatility with Excel.

Your student project compares two samples, not necessarily the two baseball teams or two Leagues. Review the cell formulas in the illustrative worksheet, and apply the same logic to your student project.}

Jacob: How do we compare National vs American Leagues for the *F* test?

Rachel: Create a data base for all teams in both Leagues. Use a dummy variable D = 0 for National League teams and D = 1 for American League teams. Place the dummy variable in Column B.

Illustration: The data sets on the NEAS web site include the League. Suppose the league name is in Column C, and the first row of figures is Row 11. Use an Excel *IF* statement as a formula in Column H. In Cell H11, write

=IF(C11="American", 1,0).

If your data has “A” for American League and “N” for National League, write

=IF(C11="A", 1,0).

Copy the formula in Cell H12 to all other cells in Column H.

Alternatively, sort the data by League and autofill with 0’s and 1’s – a low-tech solution that requires no knowledge of Excel functions or VBA.

Take heed: It is worth learning Excel’s *IF* statement. The project templates that use dummy variables are easy to implement with Excel’s *IF* statement.

Jacob: How many regression equations do we use?

Rachel: Use three regression equations.

- First examine the ordinary least squares estimators for each League separately.
- The third regression equation uses both leagues together.

Jacob: An *F* test requires two regression equations of the same form. What if the optimal regression equation has 5 independent variables for the National League and 6 independent variables for the American League? Do we use 5 or 6 independent variables for the equation with the dummy variable?

Rachel: Take your pick. In most cases, the result is the same.

- For a real statistical study, we might use 6 years.
- For the student project, use 5 years, to keep the work simple.

Jacob: Suppose we use 5 independent variables as the optimal regression equation, giving six explanatory variables. With the dummy variable, do we have 7 explanatory variables?

Rachel: With the dummy variable, we have 12 explanatory variables.

- The five additional slope coefficients are β_j', for j = 1, 2, 3, 4, and 5.
- The five additional independent variables are D × the won-loss record in the five prior years: all zero for one League and the five won-loss records for the other League.
- The additional intercept is α'.
- The additional constant term is D (either 0 or 1)

Jacob: How do we tell the *REGRESSION* add-in about these additional variables?

Rachel: Create columns for the additional variables. Suppose the won-loss records in the five past years are in column G, H, I, J, and K. Create 6 new columns.

- Column L = Column B (the dummy variable)
- Column M = Column B × Column G
- Column N = Column B × Column H
- Column O = Column B × Column I
- Column P = Column B × Column J
- Column Q = Column B × Column K

Create new column captions: if D = dummy variable, use

D (= dummy variable) for Column L, D-1 for Column M, D-2 for Column N, and so forth.

Jacob: How does the *REGRESSION* add-in handle the dummy variable?

Rachel: Instead of five columns of independent variables, we have 11 columns. The *REGRESSION* add-in uses these 11 columns as independent variables, plus one constant term.

Jacob: What is the null hypothesis that we test?

Rachel: We test if all six additional regression coefficients are zero.

F TEST AND OPTIMAL REGRESSION EQUATION

Jacob: Can we use the *F* statistic to select the optimal regression equation? We form regression equations for 1, 2, 3, …, 10 past years. We use *F* tests to see if 10 past years is better than 9 years, 9 past years is better than 8 years, and so forth.

Rachel: The *F* test is sometimes used for this, but it does not give the type of answer we need. Suppose we compare 3 years vs 4 years.

- The *F* test says: If the fourth past year adds no information, what is the probability of getting the observed results?
- We want to ask: If we have no assumptions about the optimal regression equation, which equation fits better?

The adjusted R^2 is the proper statistical test.