ILLUSTRATIVE WORKSHEETS: BOSTON, NEW YORK, *F* TEST

(The attached PDF file has better formatting.)

The project template explains the *F* test in detail and give many suggestions for its use.

- The illustrative worksheets show regression equations for Boston and New York (both in the American League) and an *F* test comparing them.
- The project template discusses the *F* test for American League vs National League, for different years, and for different sports.

The illustrative worksheets for Boston and New York show that the optimal regression equation has two past years for Boston and one past year for New York. We ask: "How different are Boston and New York?"

Your student project should explain what you are testing.

The null hypothesis assumes the same regression equation for both teams.

Jacob: The two teams clearly have different regression equations. The optimal equation uses two past years for Boston and one past year for New York. The ordinary least squares estimators differ. Why assume the equations are the same?

Rachel: We use 50 data points for each team. Losing percentages are stochastic. Perhaps the two teams have the same regression equation, and the differences stem from random fluctuation. We examine two possibilities:

1. In the long-run, the teams do not differ. Observed differences are temporary random fluctuations: they last for a few years and then fade out.
2. The teams have different long-run expected won-loss records, but the same relation among won-loss records of different years.

For scenario #1, neither α nor β differ for the two teams. For scenario #2, we normalize each team's long-run won-loss record to 50% before applying the *F* test.

We can use either one or two past years for the *F* test. The illustrative worksheet uses one past year. It includes the data for the second past year, and you are encouraged to run an *F* test with two past years.

Your student project may have three or four past years for the *F* test.

- A longer experience period is more likely in (American) football, which has few games per season and many players per team.
- If you need four or five years for the *F* test in baseball or basketball, check your work.

Copy the Boston and New York data for the current year and one past year (or two past years, depending on the *F* test) into a new worksheet.

Take heed: The instructions here delete some rows to simplify the worksheet. If year T+1 is copied from Year T and you delete Year T, year T+1 is missing a value.

- You can copy and paste losing percentages with *PASTE VALUES*. The new call has a value (a constant), not a formula. Deleting the initial cell has no effect on the new cell.
- If you use regular *PASTE*, you may have to retype one or two cell values.
- By using *PASTE TRANSPOSE* in the Boston worksheet, we avoided the whole problem.

The comments above may not make sense until you do the work. As you do the work, check for Excel error warnings (*VALUES! NAMES!*).

Use two regression equations:

- Restricted (constrained) equation: Forecast the won-loss record (losing percentage) from one past year (Year -1).
- Unrestricted (unconstrained) equation: Forecast the won-loss record from the past year (Year -1), a dummy variable, and the dummy variable times the past year.

The dummy variable is 0 for New York and 1 for Boston.

- Cell H11 has the formula =IF(LEFT(C11,1)="N",0,1)
- Cell H11 is copied to Cells H12:H110

If the first letter of the team name (Column C) is "N" (for New York), the dummy variable = 0; otherwise, the dummy variable = 1.

Take heed: You don't have to use an IF function. You can enter 0 in Cells H11:H60 and 1 in H61:H110.

Take heed: Instead of an *F* test with dummy variables, you can use the version of the *F* test with α' and β'.

Take heed: We have not normalized the won-loss records to 50%. New York has a better average won-loss record over the fifty years (a lower losing percentage), so the same regression equation is unlikely to work for both teams. We redo the analysis below after normalizing the won-loss records to 50%.

The Restricted Equation

The restricted equation uses one past year and no dummy variable. We are tempted to reason as follows:

> Each team might have a high R^2 if the past year is a good predictor of the current year. If the two teams differ, we expect a lower R^2, since the same regression equation won't work for both teams.

The restricted equation has a high R^2 (52.84%) vs 42.89% for New York alone and 44.60% for Boston alone. We explain the rationale.

The two teams have similar slope coefficients, but New York has a lower losing percentage (on average) in 1901-1960. We want to use the past experience. To make the intuition clear, we give an extreme illustration

Illustration: Suppose Boston had an 80% losing percentage (on average) and New York had a 20% losing percentage, and last year's losing percentage does **not** help predict the current year's losing percentage *if each team is examined alone*.

The optimal regression equation for each team alone has $\alpha = 20\%$ for New York and 80% for Boston, $\beta = 0$, and $R^2 \approx 0$.

If we examine the teams together, we place great weight on last year's losing percentage.

- If the losing percentage last year is high, the team is probably Boston ⇛ the expected losing percentage this year is 80%.
- If the losing percentage last year is low, the team is probably New York ⇛ the expected losing percentage this year is 20%.

The optimal regression depends on the stochasticity. As $\sigma \rightarrow 0$, $\beta \rightarrow 1$ and $R^2 \rightarrow 100\%$.

Take heed: The discussion forum postings for the project templates give the rationale for the statistical results. As you work through your student project, jot down the implications of the statistical results and include these in your write-up.

Take heed: Explaining the intuition for your results is a good habit, helping you understand the regression and avoid errors.

The Unrestricted Equation

The unrestricted equation uses three (2*k* + 1) explanatory variables and one intercept.

See the general discussion of restricted vs unrestricted equations.

- The R^2 increases from 52.84% to 54.88%.
- The ESS decreases from 0.47639 to 0.455791.

The R^2 always increases and the ESS always decreases. The *F* test says if the increase in R^2 and the decrease in the ESS reflects a better fit or simply more explanatory variables.

- It is *always possible* that a higher R^2 and lower ESS reflect more explanatory variables.
- The *F* test gives the *probability* that they reflect the number of explanatory variables, not the quality of the fit.

Degrees of Freedom

We determine the degrees of freedom for the numerator and denominator.

Numerator: The number of constraints is two (not one): the same α and same β. That is, the coefficient of D (the dummy variable) is zero and the coefficient of D × Year -1 is zero.

Take heed: The number of constraints is the number of independent variables + 1.

Denominator: 50 years × 2 teams = 100 data points. The unrestricted equation has four explanatory variables (including the intercept). The degrees of freedom for the denominator is 100 – 4 = 96.

Take heed: The write-up to your student project should state the degrees of freedom for the numerator and denominator.

ESS Version of *F* Test

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

The numerator is the difference in the ESS divided by its degrees of freedom:

(0.47639 – 0.455791) / 2 = 0.01030

The denominator is ESS for the unrestricted equation divided by its degrees of freedom.

0.455791 / 96 = 0.00475

The *F* statistic is 0.01030 / 0.00475 = 2.16931

Use Excel's *FDIST* built-in function to determine the *p* value for an *F* statistic of 2.16931 with (2, 96) degrees of freedom. Typing =FDIST(0.216931,2,96) gives 0.1198255.

You can also interpolate with the table of critical values for the *F* test in the textbook.

R^2 Version of *F* Test

$$F_{q,N-k} = \frac{(R_{UR}^2 - R_R^2)/q}{(1 - R_{UR}^2)/(N-k)}$$

The numerator is the difference in the R^2 divided by its degrees of freedom:

(54.88% – 52.84%) / 2 = 1.02%

The denominator is the R^2 for the unrestricted equation divided by its degrees of freedom.

(1 – 54.88%) / 96 = 0.00470

The *F* statistic is 0.01020 / 0.00470 = 2.170213

Validate Your Results

The regression output for the unrestricted equation confirms the results of the *F* test.

The dummy variable has a coefficient of 0.024993, a standard error of 0.071397, a *t* value of 0.350057, and a *p* value of 0.727063.

- The *p* value of 0.727 means that if the true intercept is the same for Boston and New York, the probability of an estimated coefficient for the dummy variable ≥ 0.024993 or ≤ –0.024993 is 72.7%. The high *p* value and low *t* value lead us not to reject the null hypothesis that the intercepts are the same for New York and Boston.

The (dummy variable × Year -1) has a coefficient of 0.015221, standard error of 0.156349, *t* value of 0.09735, and *p* value of 0.922651.

- The *p* value of 0.92265 means that if the true slope coefficient is the same for Boston and New York, the probability of an estimated coefficient for the (dummy variable × Year -1) ≥ 0.015221 or ≤ –0.015221 is 92.265%. This probability is high ⇛ we do not reject the null hypothesis that New York and Boston have the same slope coefficient.

The unrestricted equation says that the slope coefficient for Boston alone is 0.65651, and the slope coefficient for New York alone is 0.65651 + 0.015221 = 0.671731. This difference is small ⇛ it reflects random fluctuations in the data sample of 50 years for each team.

Take heed: Confirm your *F* test results by examining the regression equations for each team (or league or conference or division) separately.

- If the regression equations differ materially, the *F* test should reject the null hypothesis.
- If the equations do not differ much, the *F* test should not reject the null hypothesis.

Normalization

The *F* test above considers two possibilities: New York and Boston are the same or they have different intercepts and slopes. We might also consider a third possibility:

We might surmise that Boston and New York have different average won-loss records, but the past year's losing percentage has the same effect on the current year's for both teams.

The average losing percentage differs materially for the two teams. In 1911-1960:

- New York lost 41.0% of its games.
- Boston lost 49.9% of its games.

The slope coefficient for the two teams is almost the same. Boston's is higher by

$$0.015221 / 0.65651 = 2.32\%.$$

Normalization does not affect the slope coefficient, as shown in the unrestricted equation.

Jacob: Don’t we check if the long-run average losing percentages are the same by the standard deviations and the means?

- Suppose the standard deviation for each team is 10%.
- The standard deviation of the means is 10% / √50 = 1.41%.
- The probability that both teams have the same mean (say 45.5%) is extremely small.

Rachel: Your procedure is correct if losing percentages are uncorrelated from year to year. But suppose β = 99%, the true mean is 45.5% for both teams, and the losing percentage the first year is 40% for New York and 51% for Boston. With a 10% standard deviation for the losing percentage each year, this is a likely scenario: each team’s observed value is within half a standard deviation of the mean. With β = 99%, each team’s expected losing percentage is about the same from year to year, with a slow reversion to the mean.

We ask: “If we normalize both teams to an average won-loss record of 50%, how would the *F* test change?”

Method: Let B′ be the average losing percentage for Boston in 1910-1960 and N′ be the average losing percentage for New York in 1910-1960. The illustrative worksheet shows these averages at the bottom of the column labeled *forecast*. Give these averages names: Nyaverage and Boaverage, using Excel’s insert ⇛ name ⇛ define.

Copy Columns C, D, and E to Columns P, Q, and R.

- In Cell P10, write ”=C10”.
- Copy Cell P10 to rest of the cells in columns P, Q, and R.

Normalize the losing percentage figures to a 50% average won-loss record for both teams.

- For the losing percentages, do **not** use copy and paste; use the steps outlined here.
- Place the formula to =F11*0.5/NYaverage in Cell S11.
- Copy Cell S11 to all the cells in the *New York rows* in Columns S, T, and X.

NYaverage is a name, so it always refers to the same figure (41.0%).

Do the same for the Boston figures.

- Place the formula to =F61*0.5/BOaverage in Cell S61.
- BOaverage is a name, so it always refers to the same figure (49.9%).

We multiply the Boston figures by 50% / B′ and the New York figures by 50% / N′. Re-compute the averages to make sure they are both 50% (in column S).

- Recompute the figures for the restricted and unrestricted equations.
- Form the new *F* statistic.

For the restricted equation, the R^2 and slope coefficient decrease with normalization, and the intercept increases. See the discussion above for the rationale.

The degrees of freedom for the *F* test do not change. The difference in the ESS (or in the R^2) decreases almost to zero. The *F* statistic decreases almost to zero. We surely do not reject the null hypothesis that the regression equations for the two teams are the same except for the difference in the average won-loss record.

Take heed: A league (or division or conference) has an average won-loss record of 50%. If your student project compares two leagues, you do not have to normalize.

Take heed: The R^2 for a regression equation does not change if the figures are multiplied by a constant. The ESS changes by the square of the normalizing constant.

Jacob: Do we expect the new *F* statistic to be higher or lower?

Rachel: If teams differ in average quality (won-loss records), we expect lower *F* statistics.

- A lower *F* statistic means part of the difference in the regression equations reflects the different average won-loss records of the two teams.
- The same *F* statistic means the teams' losing percentages are not materially different.