Regression analysis Module 12: *F* test practice problems

(The attached PDF file has better formatting.)

** Exercise 12.1: *F*-Test

- RegSS is the regression sum of squares.
- RSS is the residual (error) sum of squares.
- TSS is the total sum of squares.
- n is the number of data points in the sample.
- k is the number of explanatory variables (not including the intercept).

An *F*-statistic tests the hypothesis that all the slopes (ß's) are zero.

A. What is the expression for the *F*-statistic using sums of squares?
B. What is the expression for the *F*-statistic using R^2?

Part A: The *F*-statistic using sums of squares is

$$F\text{-statistic} = (\text{RegSS} / k) \div (\text{RSS} / (n - k - 1))$$

ResSS + RSS = TSS, so some textbooks write this as

$$F\text{-statistic} = (\text{RegSS} / k) \div ((\text{TSS} - \text{RegSS}) / (n - k - 1))$$

Part B: The *F*-statistic using R^2 is

$$F\text{-statistic} = (R^2 / k) \div ((1 - R^2) / (n - k - 1))$$

R^2 = ResSS / TSS, so the expression with R^2 is the expression with TSS and RegSS after dividing numerator and denominator and TSS.

Intuition: The total sum of squares (TSS) is divided between the regression sum of squares (RegSS) that is explained by the regression equation and the residual sum of squares (RSS) that remains unexplained. If a greater percentage is explained by the regression line, R^2 is greater (ResSS is a greater percentage of TSS), the *F*-statistic is larger, and the regression is more likely to be significant.

(See Fox, Chapter 6, statistical inference, page 108)

** Exercise 12.2: Degrees of freedom of F-statistic

A regression model has *N* data points, *k* explanatory variables (ß's), and an intercept.

A. An *F*-test for the null hypothesis that **q** slopes are 0 has how many degrees of freedom in the numerator?
B. This *F*-test has how many degrees of freedom in the denominator?

Part A: The F-test says: "How much additional predictive power does the model under review have compared to what we would otherwise use, as a ratio to the total predictive power of the model under review?" Each part of this ratio is adjusted for the degrees of freedom.

The degrees of freedom in the numerator adjusts for the extra predictive power of the model under review stemming from additional explanatory variables. If the model under review has one extra explanatory variable, it predicts better even if this extra explanatory variable has no actual correlation with the response variable. The degrees of freedom is the number of extra explanatory variables, or **q**.

If the F-test has a *p*-value of P% with *q* degrees of freedom in the numerator, its *p*-value is more than P% with *q+1* degrees of freedom in the numerator. A higher *p*-value means that it is more likely that the observed increase in predictive power reflects the spurious effects of additional explanatory variables.

Part B: The degrees of freedom for the model under review is N – *k* – 1; this is the degrees of freedom in the denominator of the *F*-ratio. As N increases but no other parameters change, the additional predictive power of the model under review is less likely to be spurious (more likely to be real), so the *p*-value decreases

** Exercise 12.3: *F* test

A linear regression $Y_j = \alpha + \beta \times X_j + \epsilon_j$ with 5 observations has an estimated $\sigma^2_\varepsilon = 1.4333$ and an *F* value of 10.1628.

A. What is the residual sum of squares (RSS) of the regression?
B. What is the regression sum of squares (RegSS)?
C. What is the R^2 of the regression?
D. What is the absolute value of the correlation between the explanatory variable and the response variable?
E. If the ordinary least squares estimator of β is 1.7, what is its standard error?

Part A: The regression equation has one intercept, one explanatory variable, and five observations, so it has N – k – 1 = 5 – 1 – 1 = 3 degrees of freedom. The σ^2_ε = RSS / degrees of freedom ⇛

$$\text{RSS} = \sigma^2_\varepsilon \times \text{degrees of freedom} = 1.4333 \times 3 = 4.300.$$

Part B: The *F* value = the regression sum of squares (RegSS / k) / (RSS / N – k – 1) = (RegSS / 1) / (4.3 / 3)

$$⇛ \text{RegSS} = (4.3 / 3) \times 20.1628 = 28.900.$$

Part C: The R^2 of the regression is the regression sum of squares divided by the total sum of squares. The total sum of squares TSS = ResSS + RSS = 28.9 + 4.3 = 33.2, so the

$$R^2 = 28.9 / 33.2 = 0.87048.$$

Part D: The absolute value of the correlation is the square root of the R^2:

$$\sqrt{0.87048} = 0.9330.$$

Part E: The *t* value for B, the ordinary least squares estimator for β, is the square root of the *F* value:

$$t \text{ value} = \sqrt{20.1628} = 4.4903$$

This *t* value is the ordinary least squares estimator for β divided by its standard error ⇛ the standard error =

$$\text{standard error} = 1.7 / 4.4903 = 0.3786.$$