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 (Fox, section 6.2.2):

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

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: Regression statistics: estimators, sum of squares, R^2, *t* values, *F* tests

Regression statistics may have units of the explanatory variables, the response variable, both, or neither.

A. What regression statistics are unit-less?
B. What regression statistics depend on the units of measurement for the response variable?
C. What regression statistics depend also on the units of measurement for the explanatory variable?
D. What regression statistics depend on the degrees of freedom?

Part A: The R^2, the adjusted R^2, the correlation ρ, *t* values, *p* values, and *F* ratios are unit-less. These statistics measure goodness-of-fit. They are percentages (R^2, adjusted R^2, correlation ρ) or quantiles (*p* values, *F* ratios) of distributions.

Units of measurement do not affect these regression statistics. Changing explanatory variables or response variables to a different scale does not affect the goodness-of-fit.

Illustration: Suppose the R^2 for a regression analysis with the response variable measured in meters is 50%. If the response variable is measured in centimeters or kilometers, the R^2 stays 50%.

The R^2, adjusted R^2, correlation ρ, *p* values, and *F* ratios are equivalent: if the number of observations and the degrees of freedom are known, each can be converted into the others. For simple linear regression with one explanatory variable, the *t* value is also an equivalent measure,

Some final exam problems convert these regression statistics into one another. For linear regression, the *F*-ratio is the square of the *t* value and is a function of the R^2. The correlation ρ is the square root of the R^2, and the adjusted R^2 is a function of the R^2. The *p* value is too complex to be derived by pencil and paper, but it is easily read from a table of the *t* distribution or the *F* distribution.

Some conversions depend on the number of observations and the degrees of freedom (which depend also on the number of explanatory variables or the number of constraints).

Part B: The total sum of squares TSS, regression sum of squares RegSS, and residual sum of squares RSS depend on the square of the units of measurement of the response variable. TSS, RegSS, and RSS use the deviations of the response variable from its mean, so shifts in the units of measurement do not affect them.

- If the units are *k* times larger, so each response variable is divided by a constant *k*, the TSS, RegSS, and RSS are divided by k^2.
- If the units are shifted by a constant *k*, the TSS, RegSS, and RSS do not change.

Illustration: Suppose the TSS for a response variable measured in meters is 50. If the response variable is measured in centimeters (multiplied by 1%), the TSS becomes 50 ÷ 1%² = 500,000. This TSS is in units of 500,000 centimeters-squared, which equals 50 meters-squared.

Degrees Celsius and degrees Kelvin are the same units of measurement, but they have different origins. The TSS, RegSS, and RSS do not depend on the origin.

Illustration: A regression analysis measures distance from the right side a rectangular field. If the distance is measured from the left side of the field, the TSS, RegSS, and RSS do not change.

Jacob: How does the ordinary least squares estimator for the variance of the error term change?

Rachel: The estimator &r2e (s^2) is the RSS / (n - k - 1), so it has the same units of measurement as the RSS.

The population regression parameter α depends on the units of measurement for the response variable.

- If the units are *k* times larger, so the response variable is divided by a constant *k*, α is divided by *k*.
- If the units are shifted by a constant *k*, α shifts by the same constant.

Illustration: Suppose the α for a regression analysis with the response variable measured in meters is 50. If the response variable is measured in centimeters (multiplied by 1%), α becomes 50 ÷ 1% = 5,000. An α of 5,000 centimeters equals an α of 50 meters.

Jacob: Are shifts in the response variable common in statistical analyses?

Rachel: Time is measured with an arbitrary origin. Calendars have arbitrary starting points: religious calendars for Judaism, Islam, and Eastern religions have different Year 0's and different New Year's dates. Present values in actuarial and financial work use an arbitrary current date. The CPI may have a base year of 1990, 2005, or some other date.

Part C: The population regression parameter β depends also on the units of measurement for the explanatory variable. The β is a function of the x- and y-deviations from their means, so a constant shift in the units of the explanatory variable or response variable do not affect the value of β.

- If the units for the explanatory variable are *k* times larger, so the explanatory variable is multiplied by *k*, β is divided by *k*.
- If the units for the response variable are *k* times larger, so the response variable is multiplied by *k*, β is multiplied by *k*.

Illustration: A regression analysis uses meters for both the explanatory variable and the response variable, with a β of 50.

- If the explanatory variable is measured in centimeters, β becomes 0.500.
- If the response variable is measured in centimeters, β becomes 5,000.
- If both the explanatory variable and the response variable are measured in centimeters, β stays 50.

Jacob: How do the variances of the ordinary least squares estimators A and B change?

Rachel: The standard errors of A and B change in the same fashion as A and B, which change in the same fashion as α and β.

Illustration: If a change in the units of measurement causes β to be twice as large, B is also twice as large, the standard error of B is twice as large, and the variance of B is four times as large. The *t* value is B divided by its standard error, so the *t* value does not change.

Part D: The question can be interpreted two ways:

- Which statistical measures have the degrees of freedom in their computation?
- Which statistical measures have expected values that change as the degrees of freedom changes?

The computation of R^2 and of its square root ρ do not seem to use the degrees of freedom. The adjusted R^2 uses the degrees of freedom (the observations N and parameters *k*) in its computation. But as N increases, the expected value of R^2 decreases and the expected value of the adjusted R^2 does not change,

Jacob: If the computation of R^2 does not use the degrees of freedom, why does the R^2 depend on the degrees of freedom?

Rachel: Suppose the x-values are random draws from a normal distribution, so the y-values are also random draws from a normal distribution. $\sigma^2(y)$ and σ^2_ε are fixed values that do not depend on the degrees of freedom.

$R^2 = (1 – RSS) / TSS$, where TSS depends on $(N - 1) \times \sigma^2(y)$ and RSS depends on $(N-k-1) \times \sigma^2_\varepsilon$.

$(1 - R^2)$ varies with the ratio (N-k-1)/(N-1).
(1 - adjusted R^2) multiplies $(1 - R^2)$ by the ratio (N-1)/(N-k-1)

The adjusted R^2 undoes the influence of the degrees of freedom on the plain R^2. The *F*-statistic would be over-stated if we divided RSS by (n-1) instead of (n-k-1). Using the proper degrees of freedom corrects this over-statement.

Intuition: The correlation ρ and the plain R^2 are over-stated in small samples because of the low degrees of freedom. If the sample has only two points, ρ is ± 1 and $R^2 = 1$, even if the response variable is independent of the explanatory variable. The adjusted R^2 undoes this overstatement.

p values are quantiles of distributions. The degrees of freedom changes the shape of the distribution; it does not just shift or re-scale the distribution. We need *t* distributions and *F* distributions for the particular degrees of freedom (or pair of degrees of freedom) in the regression analysis,

Intuition: The adjustments for degrees of freedom eliminate the distortions in small ssmples.

Jacob: How do units of measurement affect the final exam problems?

Rachel: An exam problem may give one measure of goodness-of-fit and ask for the other measures. None of these depend on the units of measurement.

To test sums of squares (TSS, RegSS, and RSS) or the least squares estimator of σ^2_ε (S^2_E), the exam problem must give information about the units of measurement for the response variable.

To test the estimate of B (the least squares estimator of β) or the standard error of B, the exam problem must give information about the units of measurement for the explanatory variable as well.

** Exercise 12.4: Sum of squares, R^2, and *F* test

A linear regression $Y_j = \alpha + \beta_1 \times X_{1,j} + \beta_2 \times X_{2,j} + \epsilon_j$ with 35 observations has a total sum of squares (TSS) of 100 and a residual sum of squares (RSS) of 64.

A. What is the regression sum of squares (RegSS)?
B. What is the R^2 of the regression?
C. How many degrees of freedom does the omnibus *F*-statistic have in the numerator and denominator?
D. What is the residual mean square RMS?
E. What is the regression mean square RegMS?
F. What is the omnibus *F*-statistic?

Part A: RegSS = TSS - RSS = 100 – 64 = 36.

Part B: R^2 = RegSS / TSS = 36%.

Part C: The degrees of freedom in the numerator is k, the number of explanatory variables (not including the intercept) = 2 in this exercise. The degrees of freedom in the denominator = n – k – 1 = 35 – 2 – 1 = 32.

Part D: The residual mean square RMS = RSS / (n – k – 1) = 64 / 32 = 2.

Part E: The regression mean square RegMS = RegSS / k = 36 / 2 = 18.

Part F: The omnibus *F*-statistic = RegMS / RMS = 18 / 2 = 9.

Jacob: The R^2 of this regression is only 36%, but the *F*-statistic is highly significant. Is that reasonable?

Rachel: An R^2 of 36% means the regression explains 36% of the total variance. The *F*-statistic reflects the probability that we should reject the null hypothesis that the explanatory variables do not explain any of the variance of the response variable. This probability is close to zero, and the *F*-statistic is highly significant.

** Exercise 12.5: *F* test

We can derive the R^2 and residual sum of squares from the *F*-statistic, the number of observations, and the number of explanatory variables.

A linear regression $Y_j = \alpha + \beta_1 \times X_{1,j} + \beta_2 \times X_{2,j} + \epsilon_j$ with 35 observations has a total sum of squares (TSS) of 100 and an omnibus *F*-statistic of 9.

A. How many degrees of freedom does the omnibus *F*-statistic have in the numerator and denominator?
B. What is the R^2 of the regression?
C. What is the regression sum of squares (RegSS)?
D. What is the residual sum of squares (RSS)?
E. What is the regression mean square RegMS?
F. What is the residual mean square RMS?

Part A: The degrees of freedom in the numerator is k, the number of explanatory variables (not including the intercept) = 2 in this exercise. The degrees of freedom in the denominator = n – k – 1 = 35 – 2 – 1 = 32.

Part B: The omnibus *F*-statistic = $\frac{n-k-1}{k} \times \frac{R^2}{1-R^2}$, so we derive R^2 as

k × *F*-statistic × $(1 - R^2)$ = (n – k – 1) × R^2 ⇛ R^2 = q × *F*-statistic / (n – k – 1 + q × *F*-statistic) =

$$2 \times 9 / (35 - 2 - 1 + 2 \times 9) = 0.360 = 36\%$$

Part C: R^2 = RegSS / TSS = 36%, so RegSS = TSS × R^2 = 100 × 36% = 36.

Part D: RSS = TSS – RegSS = 100 – 36 = 64.

Part E: The regression mean square RegMS = RegSS / k = 36 / 2 = 18.

Part F: The residual mean square RMS = RSS / (n – k – 1) = 64 / 32 = 2.

Jacob: Can we also derive the t values in this exercise?

Rachel: If the regression equation has only one explanatory variable, its t value is the square root of the *F*-statistic. If the regression equation has more than one explanatory variable, we can not derive the t values.

** Exercise 12.6: *F* test, *t* value, R^2

A linear regression $Y_j = \alpha + \beta \times X_j + \epsilon_j$ with 5 observations has an S_E^2 (the least squares estimate of σ^2_ε) = 1.4333 and an *F* value of 20.1628.

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. What is the *t* value for B, the ordinary least squares estimator of β?
F. 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)

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

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 =

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