Module 12: Statistical inference for multiple linear regression

(The attached PDF file has better formatting.)

REGRESSION ANALYSIS PRACTICE PROBLEMS F TEST

F tests are used in the student project as well as on the final exam. The practice problems show the material needed for the final exam. (Fox uses RSS instead of ESS.)

*Question 1.1: F Tests

An insurer uses 3 independent variables and one constant term to forecast auto insurance rates. The 3 independent variables are driver age, territory, and driving record.

The insurer presumes that driving record is a significant rating variable, but driver age and territory might not be significant. The insurer uses an *F* test to determine if driver age and territory are not significant (in combination). There are 35 observations. The appropriate test statistic is

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

or

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

TSS is the total sum of squares and ESS (RSS) is the error (residual) sum of squares.

Which of the following is true?

A. ESS is greater for the restricted equation, but TSS is the same for both equations.
B. ESS is greater for the unrestricted equation, but TSS is the same for both equations.
C. ESS and TSS are both greater for the restricted equation.
D. ESS and TSS are both greater for the unrestricted equation.
E. ESS may be greater for either equation, but TSS is the same for both equations.

Answer 1.1: A

- The total sum of squares (TSS) depends only on the Y values, so it is the same for both regression equations.
- The error (residual) sum of squares is the unexplained variation. Adding an independent variable explains more of the variation. Even if the additional variable is not related to the dependent variable, it “explains” some of the variance just by sampling error.

Jacob: If the additional variable is unrelated to the dependent variable, how much more of the variance do we expect it to explain?

Rachel: This is degrees of freedom. Suppose a sample of N data points has a variance of σ^2. The total sum of squares of these N points is $(N-1)\sigma^2$. If we use an unrelated independent variable to help explain the variance, the unexplained variance remains σ^2, so the unexplained variation is $(N-2)\sigma^2$. The additional variation explained is σ^2, so the additional variance explained is $\sigma^2/(N-2)$.

Jacob: Is this true in all cases? Are these figures exact?

Rachel: These are expected figures. In any scenario, the additional variance explained may be greater or smaller. In most regressions, even seemingly unrelated variables may have some correlation. For example, if we regress the price of cars in Japan on the price of bananas in Spain, what is the expected value of β?

Jacob: These two variables are unrelated; β should be zero.

Rachel: The two items are unrelated, but both prices are affected by inflation, and inflation in Japan is related to inflation in Spain. β will be positive, even if it is close to zero.

Jacob: What if we use deflated prices for both cars and bananas?

Rachel: If the prices of cars and bananas follow random walks, we still expect a non-zero beta. Suppose we examine years 20X0 through 20X9, and 20X0 was a high price year for cars.

- If 20X0 was a high price year for bananas, 20X1 is expected to be a high price year for cars, though both prices drift back towards their means. We expect a positive β.
- If 20X0 was a low price year for bananas, 20X1 is expected to be a high price year for cars and a low price year for bananas, though both prices drift back towards their means. We expect a negative β.

Jacob: This is disconcerting. Random walks are common. If the independent variable follows a random walk, we are over-stating the significance of the regression.

Rachel: The textbook mentions this in the time series section. In the regression section, it says that the statistical tests are exact only if the classical regression assumptions are met. One of these assumptions is that the X values are *not* stochastic.

*Question 1.2: *F* Ratio and *t* Statistic

We fit 21 observations to $Y_i = \alpha + \beta \times X_i + \varepsilon_i$ $\hat{\alpha}$ with = 5.

- The means of X and Y are 1 and 3.
- The *F* statistic for testing the relation between X and Y is 1.96.

What is the *t* statistic for the ordinary least squares estimator of β?

A. –3.84
B. –1.96
C. –1.40
D. +1.40
E. +3.84

Answer 1.2: C

In the two-variable regression model, the *t* statistic is the square root of the *F* ratio: $1.96^{\frac{1}{2}} = \pm 1.400$.

The means of X and Y are 1 and 3, so $3 = 5 + \beta \times 1 \Rrightarrow \beta < 0$, so the *t* statistic is *negative*.

*Question 1.3: *F* Distribution

We estimate a multiple regression model with an *intercept*, *four independent* variables, and *55 observations*. We use the *F* statistic to test the hypothesis that *two* of the β coefficients are not both zero. What are the proper parameters for the *F* statistic?

A. F(2, 55)
B. F(3, 53)
C. F(4, 50)
D. F(2, 50)
E. F(4, 53)

Answer 1.3: D

The degrees of freedom are (q, N-k), where k is the number of explanatory variables including the constant term (the intercept).

q = 2
N = 55
k = 4 + 1 = 5
N – k = 50

*Question 1.4: *F* Distribution

Which of the following is true regarding the *F* Distribution?

A. The *F* Distribution is symmetrical and ranges in value from 0 to infinity.
B. The *F* Distribution is skewed and ranges in value from 0 to infinity.
C. The *F* Distribution is symmetrical and ranges in value from –∞ to +∞.
D. The *F* Distribution is skewed and ranges in value from –∞ to +∞.
E. The *F* Distribution is used to test hypotheses about a ordinary least squares estimator only when the variance of the estimator is known.

Answer 1.4: B

A sum of squares is non-negative, so the ratio of sums of squares is non-negative.

If all the sample points lie on a straight line (for a two dimensional model), the standard error of the regression coefficient is zero, the *t*-value is infinity, and the *F*-statistic is infinity.