Empirical exercise – Inference – F test based on the residual sum of squares

1. Aim of the exercise

The aim of this exercise is to understand using the F test based on the residual sum of squares. The empirical context is as follows. Suppose that you want to analyse the factors that determine the test scores of students in high school. You have access to data collected on test scores, student teacher ratio, percentage of English learners in a given class, among a set of other characteristics. The details of the data are given in the enclosed data definition file. You decide to use a linear regression model to explain how the student teacher ratio and the percentage of English learners affect the test scores. You use the OLS estimator to estimate the parameters of this regression model. You want to know how much the independent variables you consider matter in explaining the sample variation in the dependent variable. For this, you decide to conduct a hypothesis test of model significance using the F test based on the residual sum of squares.

2. Load the data

Load the enclosed data on test scores in mat format to MATLAB.

```
clear;
load 'M:\exerciseinferenceFrss.mat';
clearvars -except testscr str el_pct avginc;
```

3. Create the systematic component of the regression equation

Create the systematic component of this regression equation.

```
y = testscr;
N_obs = size(y,1);
x_0 = ones(N_obs,1);
X = [x_0 str el_pct];
```

4. Obtain the relevant statistics to be used to construct the F test

We want to test if the model we use is statistically significant at a desired level of significance. This is a test of the hypothesis that the explanatory variables have no effect on the the dependent variable on average, against the alternative that at least one of the coefficients has an effect: $H_0 : \beta_1 = 0, \beta_2 = 0$ against $H_1 : \beta_1 \neq 0$ and/or $\beta_2 \neq 0$.

This hypothesis can be tested using the F statistic. The F statistic takes the form $F = (SSR_r - SSR_{ur})/q)/(SSR_{ur}/(n - k - 1))$, where SSR_r and SSR_{ur} are the sum of squared residuals from the restricted and unrestricted models, respectively, q is the number of restricted parameters, and $n - k - 1$ is the number of observations less the number of model parameters.

These statistics are calculated in the enclosed function file. Study these calculations. Obtain these statistics in your MATLAB script file using the code presented at the end of the section.

```
LSS = exercisefunction(y,x_0);
RSS_restricted = LSS.RSS;
LSS = exercisefunction(y,X);
RSS_unrestricted = LSS.RSS;
F_df_restrictions = 2;
F_df_residual = N_obs-size(X,2);
```

5. Hypothesis test on the significance of the model

Given the terms of the F statistic, we can now carry out our hypothesis test. First, construct the statistic.

Second, use the built-in `finv` function to obtain a critical value from the $F_{q,n-k-1}$ distribution. The function instructs MATLAB to return the critical value that leaves a probability area of 5 percent at the upper tail of the F distribution at a given degrees of freedom.

Finally, use the `fcdf` function to obtain a p-value corresponding to the empirical value of the F statistic and the given degrees of freedom.

The empirical value of the test is well above the critical value. The p-value of the test is also virtually zero. Hence, we soundly reject the null hypothesis that the student teacher ratio and the percent of English learners play no role in explaining the variation in the test scores.

```
F = ((RSS_restricted-RSS_unrestricted)/F_df_restrictions)/(RSS_unrestricted/F_df_residua
F_critical_95 = finv(0.95,F_df_restrictions,F_df_residual);
p = fcdf(F,F_df_restrictions,F_df_residual,'upper');
```