ECO240 R Homework 2

[Due Date: May 16th (Tuesday) 2017 at 16:00]

Submit together with the signed a) contribution sheet and b) Honor code.

Task 1: Hypothesis Tests of a Single Population
Task 2: Two Population Hypothesis Tests
Task 3: Two Variable Regression Analysis

Data: *Attach your sampled data and the data used for Task 3 to the report.*
Task 1: Use midterm2017 variable after randomly sampling 50 observations.
Task 2: Use midterm2016 after randomly sampling 30 observations, and use it together with the
midterm 2017 variable sampled for Tasks 1.
Task 3: Select your own data which containing at least four variables. You should select Y(dependent
variable) and X(independent variable) meaningfully.

Task 1: Hypothesis Tests of a Single Population
For this task, we use the third column (midterm2017) of "midterm.csv" data set. Randomly sample 50 observations from the population.

Step1: Read data file

Define the midterm2017 variable as
midterm2017<-midterm[, 3]

Step2: t-test
t-tests can be conducted by using "t.test" command as follows:

t.test(midterm2017,mu=D_0, alternative = "two.sided", conf.level=0.95)

the first entry midterm2017 is the sample data, mu = μ_0 (H0: μ = 30 = D_0, for example), *alternative* can be set one of the three test types ("greater", "less" or "two.sided" for upper, lower and two-tail tests, respectively), and *conf.level* is the confidence level.

Step 3: Analysis
(a) What do you conclude [Reject/Fail to Reject H0] from the test if the significance level is 5%? You need to find t-value by R (*qt(p, df,lower.tail=FALSE/TRUE)*) as you did in the RHW1. Explain your decision.
(b) What does p-value indicate?
Conduct the following hypothesis tests by repeating Step 2 and 3.

(i)	$\alpha = 0.05$	(ii)	$\alpha = 0.05$
	$H_0: \mu = 30$		$H_0: \mu = 25$
	$H_1: \mu \neq 30$		$H_1: \mu > 25$

Task 2: Two Population Hypothesis Tests

In this task, you are asked to examine whether the mean exam score for students from 2017 was different from the mean exam score for students from 2016.

The basic syntax for two samples t-test is

t.test(x,y,mu=D₀, paired = FALSE/TRUE, var.equal = FALSE/TRUE, alternative="two.sided",conf.level = 0.95)

where x and y are two variables (in our case, they are midterm2016 and midterm2017), *mu* is the value of the presumed value difference between the population means, *paired*=TRUE if you want to test paired sample, = FALSE if you want to test the independent samples, *var.equal*=TRUE if you assume the population variances are equal and FALSE if not.

Test the following given the significance level = 5%. You have to run *t.test* first to see the value of the degrees of freedom. Then you can compute t table value. Report your decision and analysis by using t value and p value.

σ_x^2, σ_x^2 unknown, no equality assumption	σ_x^2, σ_x^2 unknown, no equality assumption
$H_0: \mu_{2017} - \mu_{2016} = 0$	$H_0: \mu_{2017} - \mu_{2016} = 5$
$H_1: \mu_{2017} - \mu_{2016} \neq 0$	$H_1: \mu_{2017} - \mu_{2016} > 5$

Task 3: Ordinary Least Squares (OLS)

In this task, you are going to use Data of your choice, run OLS regression models to determine the factors affecting selected dependent variable and analyze the outputs.

Find data set containing variables of your interest. I expect you to find **one Y (dependent variable) and at least 3 Xs (independent variables)** to answer your research question.

In general, linear regression models can be expressed as
$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n$$
where Y is a dependent variable you want to analyze (e.g. GDP, life expectancy, income, demand, GPA, housing price etc.), Xs are the independent variables that affect (or believed to be affecting) the dependence variable. β_s are the coefficients which are going to be estimated by regression analysis. This is the most basic type of econometric model.

Step 1:Decide what you want to analyze (your research question).

Step 2: Decide your estimation model. What will be your dependent / independent variables?

Step 3: Read data file to R.

Step 4: Regression analysis
In order to run a regression, we use "lm" command.
Basic syntax is as follows:

youranalysis<- lm(Y~X1+X2+X3+...+Xn, data=yourdata)

where *youranlysis*: name of the analysis which you define.
        Y: the name of the dependent variable
        X1~Xn: the names of the independent variables you select.
*NOTE: R problem is case sensitive. R and r are recognized differently. You have to enter your variable name correctly including their cases.

In order to list the regression result, you have to type

Summary(youranalysis)

Step 6: Analyze the output
Once you obtain the output of the regression analysis, it's time to analyze the result. Without this interpretation stage, your regression analysis is meaningless. Discuss the following points:

(a) What are the meaning of the estimated coefficients (Estimates)? Interpret the coefficient estimated for each independent variable.

(b) How t values are computed?

(c) t-values computed are used for hypothesis test. What kind of tests are we conducting here? Write down the null hypothesis, alternative hypothesis and the meaning of the test.

(d) Stars "***" indicate the levels of statistical significance. What does it mean if there is no star?

(e) What is the meaning of $Pr(>|t|)$?

(f) What is the meaning of R-squared?

(g) Write down your observations and policy implications/ recommendations.

WHEN YOU FINALLY FINISH YOUR HOMEWORK