# HW 2

Please do not use any built-in OLS commands (such as *lm*) in R to run regressions, use the OLS formulas instead!

**1. Univariate OLS.**   
The file called “TA\_NI.csv” contains data on 12,583 firms for year 2018. Extract the following two variables from the data set: firm size (TA) and firm profitability (NI).

1. Using the formulas provided in the lecture slides estimate the coefficients and in the following model:
2. Using the results in (a) compute   
   (i) The average value of the estimated residual   
   (ii) The correlation between and .
3. Relate the results in (b) to the Normal Equations.
4. Now alter the model as follows:  
   Estimate using the formula . This is the OLS estimate for a model with intercept restricted to be zero.
5. Compare the results in (d) to those in (a)
6. Using the results in (d) compute the average value of as well as the correlation between and .
7. Relate the results in (f) to the Normal Equations.

**2. Bias in the data.**

1. Create two correlated data vectors using the following procedure.
   1. Write an R code to draw three uniformly distributed random variables:  
       , , and   
      with 240 observations each.
   2. Define and

You will use the data you just generated to do parts (b) through (d)

1. Generate response using the following model:  
   where   
   Estimate the following two regression models:  
   Model A:   
   Model B: Record the coefficient estimates and their standard errors. Repeat the process 1,000 times: generate a new vector using the same and but drawing new values for ; use it to estimate Models A and B; record the estimates and their standard values.
2. In (b), what are your average coefficient estimates? Are they biased?
3. Compute the standard error for the coefficient estimators in (b) using your sample of 1,000 coefficient estimates
4. Compute the standard error for the coefficient estimators in (b) using your sample of 1,000 standard error estimates for the corresponding coefficients.
5. Compare the results obtained in (d) and (e) for Model A. Do the same for Model B. For each model, discuss if the results are indicative of bias.

**3. Multivariate Regression.**

The file called *PS*2*Wage.txt* contains worker wages data for 39 demographic groups. Your goal here is to study the effect of change in hourly wage (WPH) on the supply of labor measured in hours worked (HRS). The task is complicated by the fact that factors other than hourly wage affect the supply of labor. We will in particular consider two such additional factors: spouse’s annual income (ERSP) and the number of years of education (SCH).

1. Assume the labor supply depends on both the wage paid and how much the spouse makes as in the following multivariate regression model,  
   (i) Discuss what sign do you expect for and   
   (ii) Estimate the regression using OLS. Are your coefficient estimates significant and have the sign you hypothesized in (i)?   
   (iii) Provide interpretation for the intercept term . Does your estimate for make sense?
2. Delete the 19th observation and rerun the multivariate regression. What do you observe when comparing the results to those in (a)?
3. It is natural to expect the spouse’s income to be a substitute for own income. Test whether it is a perfect substitute as follows. Let and be the averages of *WPH* and *ERSP* respectively. Specify the null hypothesis as : . What is your conclusion?
4. Finally, since both HRS and WPH are endogenous, we may be interested in estimating the following multivariate regression model instead,  
   where *NEIN* is non-wage income. Interpret your coefficient estimates and test their significance. Do all coefficient signs make sense?

**4. Statistical Inference.**

Consider matrix *M* = [**1***,* **x***,* **z***,* **y**] where the first column is the unit vector, and the second through fourth columns are vectors , , and : . Suppose that the moment matrix for is

1. How many observations (rows of data) does contain?
2. What are the average values for variables and respectively?
3. Find OLS estimates for the following regression:
4. Compute the residual variance, the variance-covariance matrix of estimates, and the for the regression in (c). (Hint:
5. Test the hypothesis
6. Test the hypothesis
7. Test the hypothesis ***and*** simultaneously. Comparing (g) to (e) and (f), what do you conclude?