Simulation & Risk Analytics: Homework 1

Monte Carlo Simulation as a device to verify the properties of a model or a theory

Due Date: Friday, October 5

The assignment is due by 5:00pm. Feel free to hand it to me or email it to my NCSU address.

As you have seen already in class, the "classical" assumptions of the multiple linear regression model are the following:

- The mean of the *Y's* is accurately modeled by a linear function of the *X's*.
- The random error term, ε, is assumed to have a normal distribution with a mean of zero.
- The random error term, ε, is assumed to have a constant variance, σ^2.
- The errors are independent.
- No perfect collinearity.

Under these assumptions, it can be shown that the OLS estimator is normally distributed and unbiased. Specifically, the OLS estimators will have the following properties:

$$\sqrt{\boldsymbol{n}}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \sim \boldsymbol{N}(\mathbf{0}, \boldsymbol{\sigma}^{\mathbf{2}}(\boldsymbol{X}'\boldsymbol{X})^{-\mathbf{1}})$$

where $\widehat{\boldsymbol{\beta}}$ **is a px1 vector** with the estimated parameters, $\boldsymbol{\beta}$ **is a px1 vecto**r with the true regression parameters and ***X*** **is an NxP matrix** that holds all your regressors (the first column is a vector of ones, to represent the constant of your regression, the second column holds the data for X_1, and so on).

We make use of these properties when we construct t-tests, F-tests, etc.

All this is good, but do you really believe that this is the case in practice? I.e., does this theory really work or it is merely a theoretical device that rarely works in practice?

We can use Monte Carlo simulation, to prove or disprove the validity of this theory. Consider the case where the true regression equation is defined as follows:

$$\boldsymbol{Y_i} = -\mathbf{13} + \mathbf{0.21}\boldsymbol{X_{1,i}} - \mathbf{0.9}\boldsymbol{X_{2,i}} + \mathbf{3.45}\boldsymbol{X_{3,i}} + \boldsymbol{e_i}\text{ , where}$$

$$X_{1,i} \sim Uniform(10{,}20)$$

$$X_{2,i} \sim \chi^{10}_2$$

$$X_{3,i} \sim Normal\ (mean = 18, variance = 15)$$

$$u_i \sim Normal\ (mean = 0, variance = 100)$$

In addition, all X's are independent of each other and independent of the error term (we will relax this assumption later).

If you generate data according to the above assumption and then use them to run a linear regression of Y on the X's, will the theory prove to be correct? Will all betas be normally distributed, with the properties defined above? These questions can be answered through Monte Carlo simulation.

Follow these steps (use SAS and study the sample code for the simple OLS case):

1. Construct a dataset with 100 observations by drawing 100 random values from the respective distributions of the X's and the error.
2. Using the data you generated above, along with the "true regression equation", to simulate the values for Y.
3. You know have 100 observations for Y, X1, X2 and X3 (we don't need the error from now on)
4. Run a regression of Y on X1, X2 and X3
5. Use the t-statistic to test the hypothesis :

$$H_0 : \beta_1 = -0.9$$
$$H_a : \beta_1 \neq -0.9$$

6. Run 20,000 simulations of the above problem

Using these 20,000 results, answer the following questions:

a) Is the distribution of the betas the one suggested by the theory?

b) Using a level of α=5%, how many times do you, incorrectly, reject $H_0 : \beta_1 = -0.9$? Is this expected?

c) Repeat steps 1-6, with the only difference that the variance of the error is defined as $\sigma_u^2 = 10X_{1,i}$, i.e. make the variance of the residuals a function of X_1. This introduces the problem of heteroscedasticity. Are the betas still unbiased? Normally distributed? How about the t-statistic; how many times do you incorrectly reject the null hypothesis?

d) Repeat steps 1 through 6, but on step 4 run a regression of *Y* on X_1 and X_2 only (omit X_3). Does this omitted variable have any effect on the expected results for the β's and the t-test defined above?

e) Suppose that X_2 and X_3 have a correlation of 0.6. <u>*Make sure that you use that assumption in step 1 of your analysis*</u>. Then, follow the same steps as in question (d) above. What can you say about the result of your regression in this case? Is there any difference with part d? Why or why not?