Monte Carlo and Bootstrap Methods

G.J. LI

1 Monte Carlo Simulation

Monte Carlo simulation is to simulate artificial data by computer to verify certain statistical theories. We show how simulation methods work through the following two examples.

1.1 Asymptotic Normality of OLS Estimator

Econometric theories often tell us if certain assumptions are satisfied, the estimators in a model should possess certain properties. In the previous lecture, we have shown that $\sqrt{N}(\hat{\beta}_{OLS} - \beta)$ will converge to a normal distribution asymptotically. Apart from proving it through Central Limit Theorem, we can also verify this result by Monte Carlo simulation. That is, we can use computer to generate many artificial datasets and obtain many $\hat{\beta}_{OLS}$s and check whether they are indeed normal asymptotically. Consider the following experiment:

Experiment 1.1. 1. fix N to be a certain number

2. generate N observations of explanatory variable $x_1 \sim U(0, 10)$

3. generate N observations of dependent variable: first $\epsilon \sim t_6$ and then

$y = 1 + 2x_1 + \epsilon$. We regress y on an intercept and $x1$ and obtain an OLS slope estimate.

4. repeat step 3 S times and record the OLS slope estimate each time.

5. in the end, we can report the frequency of appearances for the OLS slope estimates by a histogram and check whether they are normally distributed by a JB test.

The following are the R codes to carry out the experiment.

```
library(tseries)
n=3 #sample size
s=1e3 #number of simulations
x1=runif(n,0,10) #just one explanatory variable
b=c(1,2) #the true value of the intercept and slope
bsim=rep(0,s)
v=6#degrees of freedom
for (i1 in 1:s){
  y=b[1]+b[2]*x1+rt(n,v)
  lmyx=lm(y~x1)
  bsim[i1]=coef(lmyx)[2]
}
print(jarque.bera.test(bsim))
hist(bsim,breaks=round(s/10),freq=FALSE)
bstd=(v/(v-2)/(var(x1)*(n-1)))^0.5
curve(dnorm(x,b[2],bstd),b[2]-3*bstd,b[2]+3*bstd,500,add=TRUE)
```

We can observe that when the sample size (N) is small, the histogram of $\hat{\beta}_{OLS}$ does not resemble normal distribution and we should quite often reject the hypothesis that it is normal. As we increase N, it becomes harder for us to reject the normality hypothesis.

1.2 Power and Size of T Test

type I: the size of t test; increase the times of test should be conducted to calculate
type II: the power of t test, the minimize the type II error by increase the size of sample.

We derived many results in the previous lecture under asymptotic contexts, i.e. when sample size tends to infinity. However, if sample size is finite, it is generally very difficult to obtain analytical results. For example, we mentioned in our last lecture that the t statistic is asymptotically valid even if the error in the linear model is not normal. It is hard to say what distribution the t statistic exactly follows for a particular sample size.

power curve

For hypothesis testing, we may be interested in the size and power of a test. The **size of a test** refers to the probability of making Type I error, which is committed by a researcher when the null hypothesis is true and is rejected. Level of significance in a test is set by a researcher and it reflects the researcher's tolerance for Type I error. For the t statistic, if it indeed follows a t distribution, given significance level equal to 5%, there will be 5% chances of making Type I error. If the t statistic does not follow t distribution, the size of the test may not be the same as the level of significance. The **power of a test** is the probability of rejecting a false null hypothesis and it is equal to one minus the probability of making Type II error, committed when the wrong hypothesis is not rejected. We want to minimize Type II error. In other words, we want our test to have high power. A consistent test rejects H_0 with probability approaching one as the sample size grows, whenever H_0

is wrong. In other words, as the sample size tends to infinity, the power of the test gets to unity. For finite sample, the power of a test depends on significance level and how wrong H_0 is. Since for finite sample, we do not know exactly what distribution the t statistic follows, it is hard to calculate the size and power of the t test by straight algebra. In such situation, we have to rely on Monte Carlo simulation to estimate them.

Consider the following experiment (see `mc1.R`):

Experiment 1.2.

- generate many artificial data sets (say S) as in Experiment 1.1;

- each data set has N observations and are generated as follows: first $\epsilon \sim t_6$ and then $y = 1 + 2x1 + \epsilon$. We regress y on an intercept and $x1$ to obtain an OLS slope estimate.

- For each data set, we calculate the t statistics for different test values to see whether we should reject H_0.

- We count the total number of artificial data sets for which we reject the test values. We then divide this number by the number of artificial data sets to obtain the size or power of the t test.

To sum up from the two examples,[1] we can describe the general procedures of a Monte Carlo experiment as below:

1. For the model under study, we set the true values for the model parameters.

[1] Computer Exercise 1 is about Monte Carlo experiments on the linear regression model.

2. Generate artificial data and carry out estimation and hypothesis testing. Save the relevant results.

3. Repeat 2 many times.

4. Summarize the results.

2 Bootstrap

Monte Carlo experiments are useful to study the estimator properties if we know or are willing to assume the data generating process (DGP). In practice, we may not be sure about how to generate our data, i.e. what distribution the data should follow. In a situation like this, it would be appropriate to use bootstrap, which sample directly from the empirical distribution of the data. Suppose we are given a sample of N observations and each observation is not correlated with the others. Though we do not know how the data are generated, we can reproduce another data set by drawing uniformly from the original data set with replacement. In other words, when we draw an observation from the original sample, each observation has equal probability of being drawn. After picking an observation, we put the observation back into the sample and draw another observation. By repeating this procedure N times we can obtain another sample with N observations, whose distribution should be the same as the original sample.

In practice, when bootstrap is applied, it is quite common to sample the estimates from the model instead of the data directly. Consider the data set `wage1.csv` associated with this lecture. If we run the following regression,

$$log(wage) = \beta_0 + \beta_1 educ + \beta_2 exp + \beta_3 tenure + \epsilon, \tag{1}$$

we can find that the estimated residuals are not normally distributed. Given that the sample size is 526, we could conjecture that Central Limit Theorem may be at work and the t statistics are still valid. We could confirm this by bootstrap. Though we are not sure about how the residuals are generated, we can sample the estimated residuals from their empirical distribution and obtain a new sample of $log(wage)$ by adding up the model fitted values in (1) and the bootstrap estimated residuals. Once a new series of $log(wage)$ is obtained, we can do another OLS and have another estimated β. By repeating these steps many times, we can construct the distribution of β and compare it to normal distribution. Note that during the bootstrap, we do not directly sample $log(wage)$ and other explanatory variables, but the estimated residuals from the linear regression model. The R code `bp.R` checks the distribution of β_2 in (1) by bootstrapping the estimated residuals from the original sample $10,000$ times. We can see that the Jarque-Bera test can not reject $\hat{\beta}_2$ is normally distributed. The variable `empv` is the estimated p-value from bootstrap to test $H_0 : \beta_2 = 0$, which is not far away from the p-value reported by R using the original sample. These suggest that the t test is indeed valid given the current sample size even though the residuals are not normally distributed.

If we regress the squares of the estimated residuals from (1) on the left hand side explanatory variables, we can find that the coefficient for *tenure* is significant, which implies heteroskedasticity should exist. Moreover, the results from Breusch-Pagan test and White test all confirm the existence of heteroskedasticity. If we regress the squares of the bootstrapped estimated residuals on the same regressors again, we can see that the coefficient for *tenure* is no longer significant. In other words, bootstrap by sampling with replacement cannot preserve heteroskedasticity in the original data. To get around the problem, we need to perform wild bootstrap. The first step of wild bootstrap is to generate u_i independently from some distribution with $E(u_i) = 0$ and $Var(u_i) = 1$ for $i = 1, 2, \ldots, N$. Next we can multiply each estimated residual $\hat{\epsilon}_i$ by u_i and use $\hat{\epsilon}_i u_i$ as the bootstrapped residual. Note that $E(\epsilon_i u_i) = E(\epsilon_i)E(u_i) = 0$ and $E(\epsilon_i^2 u_i^2) = E(\epsilon_i^2)E(u_i^2) = \sigma_i^2$. In `bp.R`, you can also find the codes for wild bootstrap. Again the simulated draws of β_2 suggest its OLS estimator should be normally distributed.

In this note, I just introduce the general steps to carry out bootstrap. For more technical details, please refer to the references associated with the R package `boot`. For the connection between bootstrap and Edgeworth expansion, please refer to Hall (1992).

Quesion: If we want to tests $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$ in (1), can we use the Wald statistic $\frac{SSR_R - SSR_U}{\widehat{\sigma^2}}$? Does this statistic follow a chi-squared distribution under the null?

References

[1] W.H. Greene, *Econometric Analysis*, 7th ed., Pearson Education, Appendix D, 2011.

[2] P. Hall, *The Bootstrap and Edgeworth Expansion*, Springer-Verlag, 1992.

[3] J.M. Wooldridge, *Introductory Econometrics (a modern approach)*, 4th ed., South Western College, Appendix C6, 2008.