Data Science.
Lectures. Week 3.
Simulation and Bootstrapping.

Filippov Max

October 14, 2022

Contents

1 Monte Carlo simulation

1.1 Basic steps of Monte Carlo simulation

- **Specify the data generating process**
  Choose data generating model. Note that data is generated with the use of random numbers, which affects the resulting sample.
- **Estimate an unknown variable or parameter**
  Using a sample generated in the first step estimate the target parameter.
- **Repeat the process N times**
  The final estimation of the parameter is the mean of N measurements.

1.2 Random number generation

Congruent linear generator is a basic pseudo random number generation algorithm. It has parameters a, b, m and seed y_0 and is defined by formula $y_{i+1} = (ay_i + b) \mod m$. a, b, m are chosen such that final sequence satisfies statistical properties of a random one, and the first k elements of the sequence are usually thrown out. In practice, more complex generators are used, but they are much heavier computationally.

2 Monte Carlo sampling error and ways to reduce it

2.1 Monte Carlo sampling error

Accuracy of simulation can be estimated with standard error of the true expected value $s/\sqrt{N}$, where s is the standard deviation of the output variables, and N is the number of trials. To increase precision by 10 times, one should take 10^2 more measurements, which is computationally heavy. Rather one should try reducing s.

2.2 Anthithetic variables

Let u_t be uniform [0,1]. Let g be the function of a distribution. Then $g(u_t)$ is the original set of random draws, and $g(1-u_t)$ is a set of anthithetic variables. The latter could be used to decrease the number of required generated random numbers in two, at the same time maintaining the same number N.

$$Corr(u_t, 1-u_t) = -1 \Rightarrow Corr(g(u_t), g(1-u_t)) < 0 \Rightarrow Error = \frac{s\sqrt{1+\rho}}{\sqrt{N}},\ \rho = Corr(g(u_t), g(1-u_t))$$

2.3 Control variables

Let x be an unknown target variable. Let y be a variable similar to x, but with known properties. Let $\hat{x}$ and $\hat{y}$ be estimate variables for x and y respectively. Then x could be estimated more precise as $x^* = \hat{x} + (y - \hat{y})$, but the upgrade from default Monte Carlo simulation is achieved only if $Var(\hat{y}) < 2Cov(\hat{x}, \hat{y})$. The inequality is derived from $Var(x^*) = Var(\hat{x} + (y - \hat{y})) = Var(\hat{x}) + Var(\hat{y}) - 2Cov(\hat{x}, \hat{y})$

2.4 Reusing sets of random numbers

It's tempting to reuse sets of random numbers to generate data, and process is likely to converge faster if doing so. However, it doesn't reduce error of estimates compared to modelling with smaller N without reusing sets of random numbers. This method could be used to increase reproducibility of the obtained results. Also used in the bootstrapping method.

3 Bootstrapping method

3.1 Basic steps of Bootstrapping method

The main idea behind bootstrapping method is drawing a smaller subsample N times and calculating estimate of the target variable as a mean of N estimates, where subsamples are of the same size and drawn with repetition. The key difference from Monte Carlo method is that data used is historical, not generated.

3.2 Problems and limitations

The bootstrapping is based on the assumption that all subsamples have the same distribution, which can be untrue. Outliers in the data might force to take a large number of replication in order to mitigate estimate shifts. If data distribution changes significantly across the set of data, samples could be biased and also short-term changes are not captured. In case of autocorrelation present in the data, samples should be drawn by blocks of sequential data.

4 Pros and cons of simulation approaches

4.1 Disadvantages

- High computation costs
- Estimates are imprecise
- Results are difficult to replicate
- Results are highly dependent on model assumptions

4.2 Advantages

- Allows to model situations not present in historical data, but theoretically possible.