Simulation

Christoph Hanck

Summer 2023

Simulation

Definition: Simulation

Simulation refers to the process of using a random process that that can be controlled (usually, a DGP) to produce data that can be evaluated with a given method.

- Using a controlled random process, we can simulate data from a true DGP of our choice
- Simulation allows to choose whether there is a back door, or a collider, or whether the error terms are correlated, or whether there is common support

Anatomy of simulation

The typical simulation study

A typical statistical simulation has only a few parts:

- A procedure for **generating data** (an implementation of a DGP)
- An **estimation** step
- A way of **iterating** steps 1-2 many times
- A way of **storing estimates** from each iteration
- An **analysis of the distribution of estimates** across iterations

Anatomy of simulation

Example: Flipping a coin 100 times and estimating probability of heads

Steps:

Decide upon the DGP. For example a coin that shows heads with probability 75%.

1. Use the computer to produce a sample of 100 coin flips with 75% probability of heads.
2. Apply the estimator to the sample of 100 coin flips. In other words, calculate the proportion of heads in that sample.
3. Store the estimate
4. Iterate steps 1-3
5. Examine the distribution of heads

Use of simulation

- Simulation can **identify bad estimators**: in the coin tossing example, simulation can be used to compare the proportion of heads estimator (the sample mean) to any other estimator.
- Simulation allows to **infer an estimator's sampling distribution**: it can help to figure out the degree of sampling variation without having to write a proof.
- Simulation is crucial to **check robustness against violations of the assumptions of an estimator**, and how sensitive its estimates are to changes in the DGP.

  If a true model has a second-order polynomial for X, but we estimate a linear regression without that polynomial, the results will be wrong. Simulation can help to know *how wrong*.

Use of simulation

Causal effects

In the context of causal inference, simulation can determine if a an estimator provides, *on average*, the true causal effect.

We may check whether the mean of the effect estimates is near the truth and their distribution is not too wide.

Situations where simulation works well

Simulation is a great tool for three main things:

- Trying out new estimators
- Comparing different estimators (*horse race*)
- Seeing what needs to be broken in order to make an estimator stop working.

Horse race

Example: Choices in matching process

- Different choices can be made in a matching. The most important choice is between **reducing bias** and **improving precision.**
- Allowing a wider caliper brings in more, but worse matches, increasing bias but also making the sampling distribution narrower.
- It might be easier if the amount of bias reduction is not very high, but the precision savings are huge. To know such a thing, **horse race simulation is useful.**

Breaking things

Example: Linear regression

- If we want a linear regression coefficient on X to be an unbiased estimate of the causal effect of X, we need to assume that X is unrelated to the error term (*no open back doors*).
- In social sciences there are always some thing we cannot measure or control for which is unfortunately related to both X and the outcome.
- Simulation can be used to probe these kinds of issues using a simulation function that is flexible enough to gauge the effect of different kinds of assumption violations or degrees of those violations.

Power analysis

Definition: Power analysis

Power analysis is the process of trying to figure out if the amount of information present in the data is enough. While there are calculators and analytical methods out there for performing power analysis, power analysis is very commonly done by simulation, especially for non-experimental studies.

Power analysis

What can go wrong with statistical analysis?

- Small effects are very hard to identify.
- Most statistical analyses involve looking at variation. If there is little variation in the data, the analysis will be tricky.
- If the data is very noisy, then analysis becomes very difficult.
- If the standards of evidence are set very high, then a lot of good evidence can get ignored and more evidence is required to make up for that.

Why is it a good idea to do power analysis?

All of the above problems can be fixed by increasing the **sample size.** Power analysis is the way of figuring out exactly how much data we need to satisfy reliability standards.

Power analysis

When can power analysis be used?

Power Analysis can be mainly used in three situations:

- If the analyst has a impression that an effect is probably not that central or important or has a small effect on a part of a system where a lot of other "stuff" is going on, a power analysis can be performed.
- If the variation of an effect over a group is of interest, then a power analysis is a good idea. Finding differences between groups in an effect takes more data to find than the effect itself.
- In a randomized experiment a power analysis is most importance as the sample size can be controlled

Power analysis

What power analysis does

Let X represents treatment and Y represents the outcome and consider

1. the **true effect size** (coefficient in a regression, a correlation, etc.),
2. the **amount of variation in the treatment** ($Var(X)$),
3. the **amount of other variation in** Y (either the variation from the residual after explaining Y with X, or just $\mathrm{Var}(Y)$),
4. **statistical precision** (the standard error of the estimate or statistical power, i.e., the true-positive rate),
5. the **sample size**.

A power analysis holds four things constant and allows conclusion about the fifth.

Statistical precision and statistical power

Definition: Statistical precision and statistical power

- **Statistical precision** in power analysis is measured by a target level of statistical power (hence the name *power analysis*)
- **Statistical power**, also known as the true-positive rate, is a concept specific to hypothesis testing:
  It is the proportion of times we correctly reject the false null

Statistical precision and statistical power

Some notes on statitical power

- Statistical power depends on the test:

  For a hypothesis test at the 95% confidence level, it is more likely to reject the null (and thus will have higher statistical power) than for a hypothesis test at the 99% confidence level.

- There is a tradeoff: the more careful one is about false positives, the more likely one will get a false negative.

- Power analyses do not have to be run with statistical power in mind

Power analysis

Requirements

- In order to do power analysis, one must fill in the values for four of the five pieces, so that power analysis can tell the fifth one
- A prior research needs to be done to fill in the gaps
- If power research is not at all possible, then it can be substituted with an educated guess
- A measurement of standards (statistical power) needs to be set

Simulation with existing data: the bootstrap

Definition: Bootstrap

Bootstrapping is the process of repeating an analysis many times after **resampling with replacement**.

- The resampling process mimics the *sampling distribution* in the actual data.
- On average, the means of the variables will match the original means, as will the variances, the correlations between the variables, and so on.
- In this way, we can run a simulation using actually-existing data.

The bootstrap

Limitations

- Unlike with a simulation where we sample data from a *known* DGP, in bootstrapping the true DGP is *unknown*
- As a simulation tool, the bootstrap is limited to cases where the question does not rely on comparing the results to the truth
- The bootstrap is often is to estimate *standard errors*:

  By mimicking the whole sampling process, strange interdependencies between the variables are automatically simulated, allowing any oddities in the true sampling distribution to creep into the bootstrap-simulated sampling distribution.

The bootstrap

Algorithm for boostrapping a standard error

1. Start with a data set with N observations
2. Randomly sample N observations from that data set *with replacement*, allowing the same observation to be drawn more than once. This is called a **bootstrap sample.**
3. Compute the statistic of interest using the bootstrap sample. This is called a **bootstrapped statistic**.
4. Iteration: repeat steps 1-3 multiple times
5. Examine the distribution of the simulated statistics. The **standard deviation** of that distribution is the **standard error of the estimate.**

The bootstrap

Asumptions

Bootstrapping comes with its own set of assumptions to provide good estimates of the standard error.

- A reasonably large sample size is required
- The data should reasonably be well behaved. If the variables follow extremely highly skewed distributions, bootstrap will not work well.
- Care must be taken when accounting for how the observations might be related to each other

The bootstrap

Variants

There are many types of bootstrap. The most important ones are:

- **The cluster/block bootstrap**
- **Residual resampling**
- **The wild bootstrap**

The bootstrap

The cluster/block bootstrap

- The cluster bootstrap is like the regular bootstrap, except instead of resampling individual observations, clusters of observations also known as **blocks** are resampled
- It is mainly used where the data is hierarchical in nature or panel data. Hierarchical data is when observations are nested within groups. In panel data the same individual (person, firm, etc.) is observed multiple times.

The bootstrap

The cluster/block bootstrap

- Another place where the cluster/block bootstrap pops up is in the context of time series data, where the same value is measured across many time periods. This often means there is autocorrelation.
- When applying a bootstrap to time series data, the time series is first divided into blocks of continuous time. The blocks themselves are often determined by one of several different available optimal-block-length algorithms

Residual Resampling

Instead of resampling observations, in residual resampling the residuals are resampled.

The bootstrap

Residual Resampling: Algorithm

1. Perform the analysis using the original data (e.g., OLS regression)
2. Use the analysis to get a predicted outcome value $\widehat{Y}$ and also the residual $\hat{r}$ which is the difference between the actual outcome and the predicted outcome,

$$r = Y - \widehat{Y}.$$

3. Perform bootstrap resampling to get resampled residuals r_b .
4. Create a new outcome variable by adding the residual to the actual outcome, $Y_b = Y + r_b$.
5. Estimate the model as normal using Y_b.
6. Store the result and iterate steps 3 to 5.

The bootstrap

Residual resampling

The idea of the algorithm is that the predictors in the model never change with this process. Hence, the residual bootstrap mimics whatever kind of interdependence they have.

Caution:

- A downsie of residual resampling is that it does not perform well if the errors are in any way related to the predictors.
- This is a *stronger* assumption than even the OLS assumption that errors are on average unrelated to predictors!

The bootstrap

The wild bootstrap

- The wild bootstrap is popular because it works under heteroskedasticity — even when the exact form of the heteroskedasticity is unknown!
- It also performs well when the data is clustered and even when the clusters are of different size
- For a wild bootstrap, the general outline of residual resampling is followed. Except that the residuals are not resampled. Each residual needs to be lined up with its original observation each time the bootstrap data is created.