

Statistical inferences

Lucie Lu

May 13, 2018

Checking unbiasedness and consistency of the estimates in the OLS model

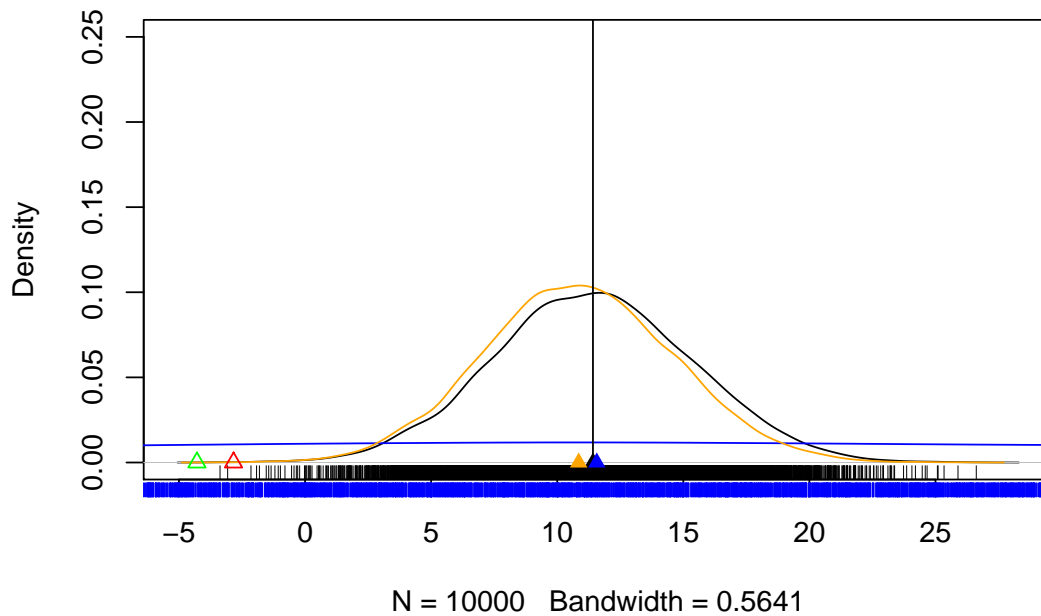
% latex table generated in R 3.4.3 by xtable 1.8-2 package % Sun May 13 21:32:25 2018

	bestATE	unbiasedATE	bestATE2
trueATE	11.42	11.42	11.42
sampmeans	11.42	11.57	10.85
bias	0.00	0.15	0.57
sd	3.96	31.82	3.78
MSE	15.64	1012.36	14.58

Table 1: Simulation results from different estimates for OLS model

Warning in rug(estdists["unbiasedATE",], col = "blue", line = 0.5): some values will be clipped

Simulation results from estimates for different OLS models



The *bestATE* is the estimator in a *lm* function with all the relevant covariates in the model. The *unbiasedATE* is the estimator in a *lm* function with no covariates in the model at all. The *bestATE2* is the estimator in a residual-based function.

To assess biasedness, I compare sample means of the estimators with the true means I created in the simulation test. In a simple way, I can check the third row bias (denoted as the absolute positive values: the differences between trueATE and the means of the estimates). All of the three estimators are close to the true mean (11.416) I created in the simulation test. This suggests all three of them are (pretty much) unbiased. In fact,

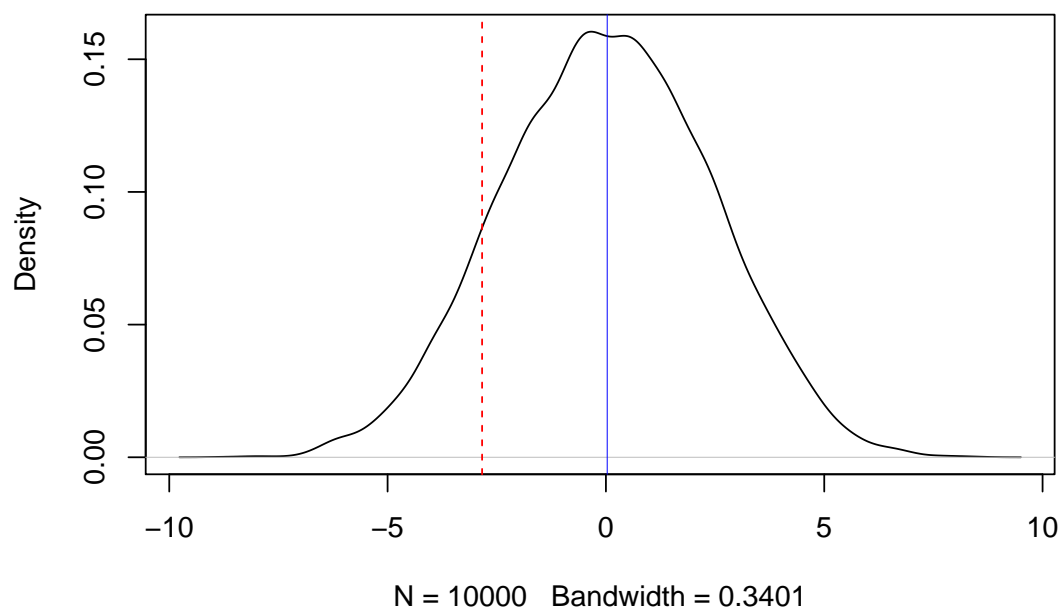
the *bestATE* has the lowest bias out of the three. Its absolute distance to the true mean is the smallest one, with bias equals to 0.0025. The *bestATE2* is slightly biased (with bias 0.5661).

To assess consistency or efficiency, the MSEs for *bestATE* and *bestATE2* are relatively the same, but the *bestATE2* behave slightly better ($14.575 < 15.6409$). The estimators *bestATE* and *bestATE2* are efficient, compared to the *unbiasedATE*. The standard error for *unbiasedATE* is very high (31.8188), suggesting that this estimate is very inefficient and inconsistent. It does not converge to the true mean at all. From this simulation test, because the *bestATE* has the lowest bias and is the most efficient one, this estimate is preferred.

P-values from the permutation test

P-value tells us how likely I can get the observed treatment effect from your experiment from the perspective of the no treatment effect hypothesis. After I have done the hypothetical experiment, I would do a hypothesis testing. The worrisome is the Fisher's sharp null hypothesis: there is a possibility of no effect for all the units in this hypothetical experiment. We just observe the differences-in-means by chance. My null hypothesis is "there is no treatment effect between the treated and control groups for each country." In other words, there are no differences in the tariff rates between democratic and non-democratic countries.

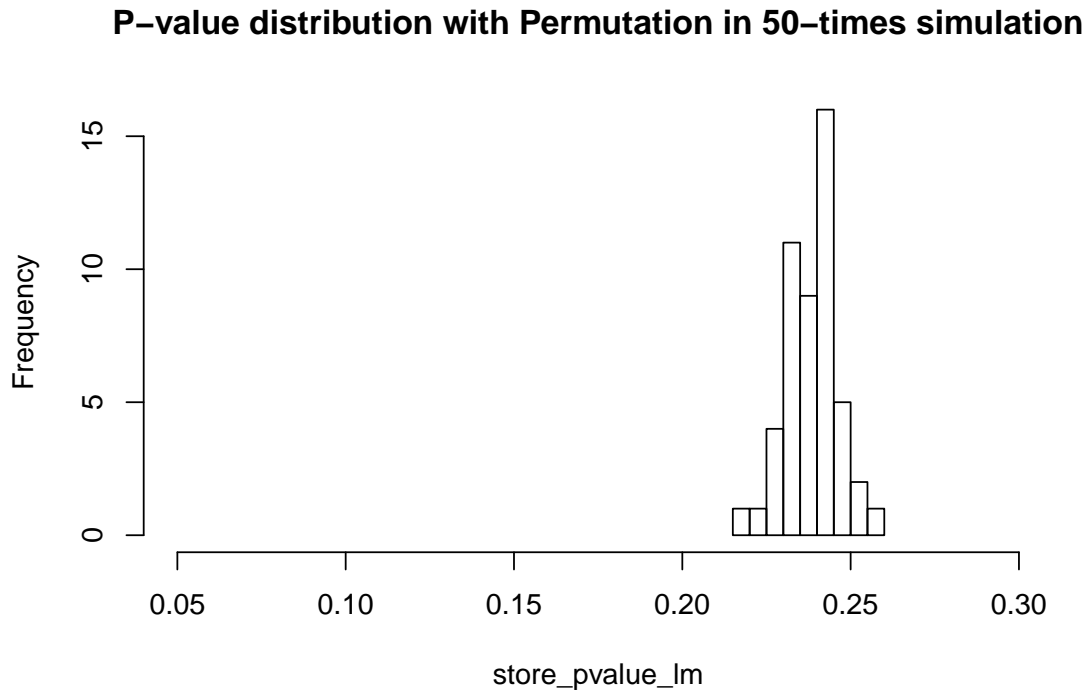
Sampling distribution under null hypothesis and Observed Statistic



First, I use a test statistic to summarize my observed data from the experiment: -2.8357 is our test statistic of mean difference in tariff rates. Second, I set the null hypothesis of no effects: I create a new experiment and shuffle the labels of country. Here, I break the relationships existing in the data structure to create an experiment of no effect. Then, by using the computing power, I replicate the experiment in the testing on the computer as if I run the experiments 10000 times. Then I observe the differences-in-means 0.0295 under the null hypothesis, and we compare how likely it is to get the differences-in-means greater than or equal to the observed data. This probability is the p-value 0.235. It means we have 0.235 (around 1 in 5 replications of the no effect experiment) to produce the values as large as or greater than the estimators in the *lm* function. The p-value here is the probability that a value as extreme or more extreme will be observed under the null hypothesis.

We use permutation test to obtain p-value where the relationship between the treatment and outcome variables is shuffled and the test statistic is calculated. The key advantage of this test does not rely on any assumptions. In the canned `lm` function, the standard assumption that the statistic follows a t-distribution gives a p-value of 0.23 which is in quite good agreement with the p-value obtained in the permutation test 0.235. But we would not necessarily know beforehand that they would agree. The following figure shows the null distribution obtained from using the data itself is close to a t-distribution. This can explain why the p-value from the CLT+IID justified test and the p-value from the permutation test is similar.

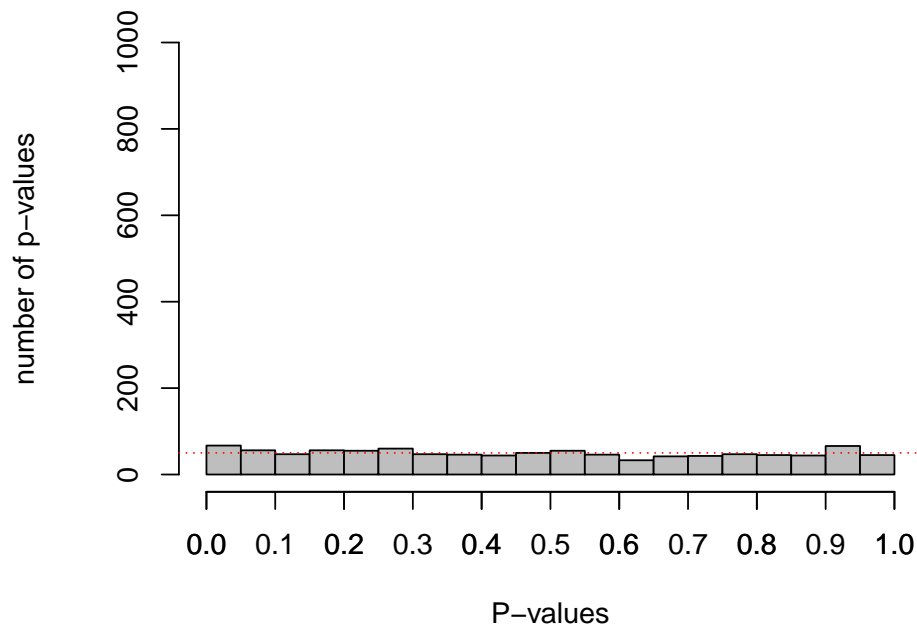
[1] 0.2268



After I calculated the p-value 0.235 from a permutation test, I replicated this process for 50 times and calculated 50 different p-values generated from the same process. From this histogram, we can see that the p-values are distributed around from 0.22 to 0.26.

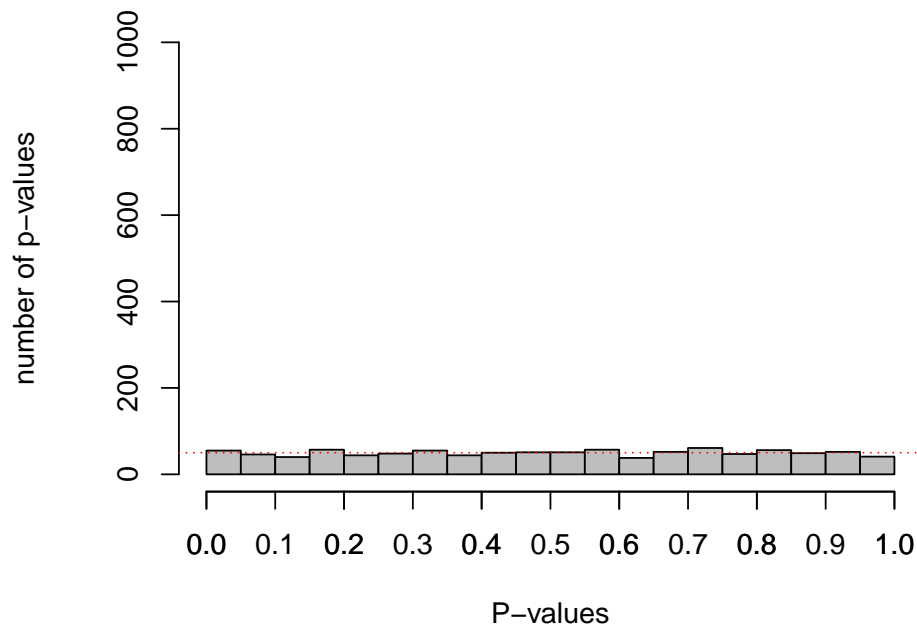
To check the error rate of the *lm function*, in the error rate test, I create a null effect knowing that my null hypothesis is true. If the false positive rate is 0.05, this means 5 out of 100 times, the test falsely reject the nulls (knowing the null is true but I still reject it). If the false positive rate is close to 0.05, it means the test fulfills its promises. The false positive rate is 0.046. If we run it a couple of times, the false positive rates are slightly different, but they are around 0.05. The p-value from the built-in `lm` function has a similar false positive rate to the nominal false positive rate (0.05).

P-value Distribution under Null Effect in the Simulation in Im Canned Fi



In this plot, we know that when there is no true effect, p-values are what is called 'uniformly distributed under the null'. The p-value distribution is basically flat. Every p-value is equally likely when the null hypothesis is true, and every bar in the graph will contain 5% of all the p-values (as indicated by the dotted red line). The first bar is the false positive rate, which is slightly higher than but it is very close to 0.05.

P-value distribution under null effect in the simulation in Imrob



I also followed the same procedures to calculate the p-value and false positive rates of the *lm rob* function. I summarize the results in the following table.

% latex table generated in R 3.4.3 by xtable 1.8-2 package % Sun May 13 21:44:13 2018

	bestATE	unbiasedATE	bestATE2
trueATE	11.42	11.42	11.42
sampmeans	11.42	11.57	10.85
bias	0.00	0.15	0.57
sd	3.96	31.82	3.78
MSE	15.64	1012.36	14.58

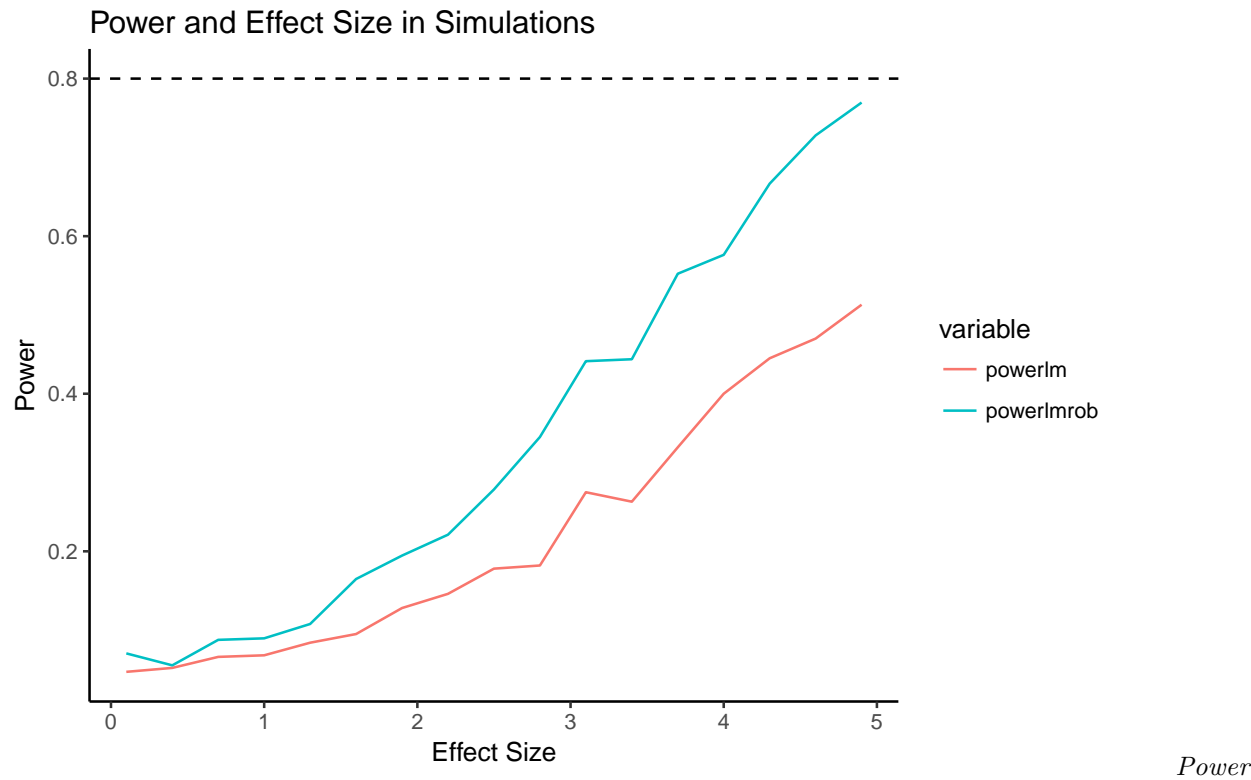
Table 2: P-values obtained from simulation, t-distribution and their erro rates

Power and effect size

When our study has effects, we hope that our test has the power to detect the true effect when the null hypothesis is false. Increasing the power of the test requires bigger sample sizes, or studying larger effects. Here, my sample size is 81, and I want to check which test (from *lm* or *lmrob*) has higher power for different effect sizes.

% latex table generated in R 3.4.3 by xtable 1.8-2 package % Sun May 13 21:47:14 2018

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
effectsize	0.10	0.40	0.70	1.00	1.30	1.60	1.90	2.20	2.50	2.80	3.10	3.40	3.70	4.00	4.30
somepower	0.05	0.05	0.07	0.07	0.08	0.10	0.13	0.15	0.18	0.18	0.28	0.26	0.33	0.40	0.44
somepower_lmrob	0.07	0.06	0.09	0.09	0.11	0.16	0.19	0.22	0.28	0.35	0.44	0.44	0.55	0.58	0.67



We can use simulations to estimate the statistical power of a model. The statistical power is the probability of observing a statistically significant result, if there is a true effect. When there is an effect, I hope that my statistical test is able to detect it. This denotes to high power in my study.

Cohen describes effect size as “the degree to which the null hypothesis is false.” In this simulation test, I generate different hypothetical effect sizes (from 0.1 to 5), and I calculate the number of p-values that are lower than 0.05 (“reject the null”) when I know there is a true effect (the null is false). When the effect size increases, the powers in both functions also increase.

For a given sample size, the *lmrob* model has larger statistical power given an effect size. As effect size increases, the power of the *lmrob* model is also increasing faster than that of the *lm* model. This is probably due to a relatively small sample size in this study (81 countries). To achieve an ideal 80% statistical power, the *lmrob* model requires an effect size larger than 5. 80% statistical power essentially means when there is a true effect, there is 80 percent that I will observe a significant effect. For this *lm* model, I need a bigger effect size to achieve the same level of power as *lmrob* model requires.