ST509 Final Project

2015150040 통계학과 배현주

Purpose

To compare some methods' performance of various interval estimates of β, this project covers mainly least square method using Majorize-Minimization (Minorize-Maximization) algorithm, bootstrap using Householder transformation regression, and Bayesian method using Gibbs sampling. All methods are under mean regression to compare which methods are still adequate to the situation where mean regression or simple linear regression would have worse performance. The purpose of this final project is to find the most efficient method under the strict condition model.

Simulation Set-up

The simulation set-up is different from assumptions required in simple linear regression. To estimate β of simple linear regression mode under nonlinear situation where x and y are in quadratic relationship, this project sets up 3 circumstances.

$$1)\ y_i = \beta_0 + \beta_1 {x_i}^2 + \epsilon_i \ for\ i = 1, \dots, n$$
$$2)\ x_i \sim N(10, 10^2)\ for\ i = 1, \dots, n$$
$$3)\ {}^{1}/_{\epsilon_i} \sim Gamma(1, 0.1) for\ i = 1, \dots, n$$

The number of the data is 100 and true values of β_0 and β_1 are 3, 10 and the true value of ϵ_i is 1. Also, the number of simulation is 10000 to compare the methods more precisely. The significance level is 0.05 which is usually used.

Performance Measure

Two performances measures will be used. First, by using MSE (Mean Squared Error), model's fitness could be summarized into numerical values. Second, by using confidence/credibility interval (as CI), the length of CI in each simulation would be calculated, and the mean and the standard deviation of CI would be used for the performance measure. In the conclusion part, the estimate values, MSE, and the mean and standard deviation of confidence/credibility interval length would be compared.

Comparison

Before comparing 3 methods, there would be 2 ways to represent the estimates' confidence interval. In Gibbs Sampling, the expression 'credibility interval' is better than confidence interval because Bayesian approaches are different from frequentist approaches in parameters. The former assigns a probability distribution to parameters and the credibility interval is a range of values on the posterior probability distribution.

Table 1. Comparison of 3 methods

	LS using MM algorithm	**Bootstrap using Householder Regression (bias corrected)**	**Bayesian using Gibbs Sampling**
Estimate of β_1 (true value: $\beta_1 = 10$)	10.006222	9.999955	9.999766
MSE	107.5751	101.0585	101.0520
Confidence/Credibility Interval Length mean	1.705593e-02	1.479853e-02	6.400877e-05
Confidence/Credibility Interval Length standard deviation	0.003524428	0.005877389	0.006121028

In the table 1, all methods estimate β_1 very closely. The difference between the estimates and true value is within 0.01. Bootstrap and Bayesian has less difference than MM method because the bootstrap and Bayesian using Gibbs sampling are relatively affected by the assumptions of simple linear regression. In bootstrap, the bias caused by the simple linear regression conditions is corrected by bootstrapping. In Bayesian method, importance of prior probability is very small which means that the variance of the prior probability distribution is very large because of the 3rd condition in simulation set-up.

In addition to the estimates, the Bayesian's MSE is the smallest among those 3 methods. Because the relation between x and y are quadratic, simple linear regression is not as good as in the situation where x and y are in linear relation. Despite of it, Bayesian has the smallest MSE among 3 methods because Bayesian uses prior-information which contains less information and flat distribution while other 2 methods are performed on the assumption of simple linear regression which is not proper in the dataset.

As MSE, the mean interval length is smallest in Bayesian. The standard deviation of interval length is not small (least squares method using MM algorithm's is the smallest which is 0.003524428) but the mean interval length is much smaller than others and the standard deviation values are close to zero, so the standard deviation of the interval length is just for reference. As the mean of the interval length is similar to zero, this means that the estimates' variance is also similar to zero.

Conclusion

In this final project, the Bayesian method is better than other two methods which represents for frequentists' methods. The reason Bayesian has the best performance in this final project is that Bayesian method is relatively free from the model assumption. Bayesian estimates the coefficients using prior distribution and likelihood, not the simple linear regression's conditions. First, the prior

distribution of the parameters is similar to the dataset's distribution. In addition to this similarity, the prior distributions' variance is large $(1 \div (0.1)^2 = 100)$ so that the effectiveness of the prior distribution is insignificant.

For the above reasons, Bayesian's estimate is the closest. However, Bootstrap method using Regression should be taken into consideration. Although the Bayesian method is better than the other two methods (Least Square methods, Bootstrap), bias-corrected bootstrap method is comparable to Bayesian method. By using householder transformation regression, the estimates would be biased but this bias problem has been solved correcting bias in the process of bootstrap.

Therefore, the Bayesian method has great performance due to the prior distribution and the bootstrap method is as good as the Bayesian method. To use Bayesian method more efficiently, it is important to set-up the prior distribution less informative and close to original dataset distribution.