**Introduction**

Linear model is a very powerful statistical analysis tool, and model selection is a very important topic of data science community. Many criterion are proposed to control the right amount of complexity to choose the best model. In this study, I will find out which criterion is the best to do model selection.

**Criterion used**

R2: where p is the number of predictors

R2 adjusted:

Mallow Cp: where P is the total number of predictors (full model)

PRESS: for linear regression where where ei is residual and hi is the diagonal element of hat matrix

AIC:

BIC:

**Methodology**

For this simulation, I will generate 10 different variables predictors. The variables are generated using multivariate normal distribution function. 100 data points are generated for each variable. Then the target variable, Y, will be generated following this distribution:

So Y chooses 5 different predictors. Beta0 to beta5 coefficients are generated by uniform distribution(min=1, max=10). Errors are generated by normal distribution with mean 0 and standard deviation 5.

My intention is to test how many times the criterions choose the right variables. I also want to test how well the models that the criterions choose perform. For this reason, I split the data into training and testing part. The training part consists of 80 data points, and the testing part consists of 20 data points. After the criterions choose the model on training data, the model is then tested on the test data, and MSE is recorded. So I will still mainly focus on number of times criterions matches the right model, but MSE will be my secondary consideration

For each setting, I will do 100 simulations. A simulation will generate 10 predictors, then generate Y, split data into training and testing. Perform an extensive search on the 10 predictors of the train data to choose the model based on 6 different criterions: R2, R2adjusted, Cp, press, AIC, & BIC. The model chooses best R2 adjusted based on maximum value, and chooses best Cp, press, AIC & BIC based on minimum values.

For R2, if I choose maximum value, I will get all predictors every time, and the result will not be interesting. For this reason, I set up a function where if I cannot increase my R2 by 0.001 (I choose this value because commonly we report the R2 to 2 decimal points), I will not replace the value of R2 square. By this stopping rule, the model chosen by R2 is the model where we stop when R2 value does not improve a lot, which is the guideline in the notes.

I will perform this study in 3 settings:

**Setting 1**: all 10 predictors are uncorrelated.

**Setting 2**: all 10 predictors are correlated.

**Setting 3:** some of the predictors are uncorrelated and some of the predictors are correlated.

**Results**

**First setting:**

The simulation results are reported in the following tables

|  |  |  |
| --- | --- | --- |
| Criterions: | Number of time the criterions matches the right model | Average MSE of test set. |
| R2 | 26 | 29.2 |
| R2adj | 21 | 29.3 |
| Cp | 47 | 28.8 |
| Press | 42 | 28.7 |
| AIC | 42 | 28.9 |
| BIC | 75 | 28.0 |

**Table 1**: simulation results of the independent covariates setting

So to my surprise, the R2 chosen by the arbitrary stopping rule outperforms the R2adj. in terms of number of times it matches the right model as well as average MSE. The BIC outperforms other criterions by a large margin. Cp is the second best result. AIC and press have the same performance on the number of times it matches the right model, but press outperforms in terms of MSE by a small margin. That means that the wrong models that press choose on average are slightly better than the wrong models that AIC choose. So the performance of the criterions on this setting from best to worse are: BIC, Cp, press, AIC, R2, R2adj.

**Second setting**

I decide to test the covariates at different correlation value. The covariates this time will all have the same value of correlations. The simulations are done at 4 different values of correlations: 0.2, 0.4, 0.6, 0.8.

The results are reported at the following table. For simplicity, I just report the number of times the criterions matches the right model.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Criterions: | Correlation = 0.2 | Correlation = 0.4 | Correlation = 0.6 | Correlation = 0.8 |
| R2 | 37 | 43 | 54 | 38 |
| R2adj | 13 | 18 | 15 | 10 |
| Cp | 33 | 29 | 36 | 25 |
| Press | 30 | 29 | 37 | 22 |
| AIC | 31 | 27 | 33 | 22 |
| BIC | 66 | 53 | 53 | 38 |

**Table 2**: number of times the criterion model matches the right model at different correlation of covariates value

we can see that the BIC almost always outperforms other methods. In general, the as correlation increases, I see that the number of time the criterion finds the right model decreases but does not monotone decrease. I think its because its harder to choose the right predictors when the predictors are related.

However, I find that using an arbitrary stopping rule of 0.001 actually outperforms really well, outperforming all other criterion except BIC in all of the correlated cases. So its an indication that human judgment can still be necessary when choosing the model.

In general, BIC works best, following by the stopping rule of R2, follow by Cp, Press and AIC perform similarly, and R2adj performs the worst.

**Third setting:**

For this setting, I first generate 3 covariates with covariate 1 independent of covariate 2 and 3, but covariate 2 and 3 are dependent with correlation 0.25. Then I generate the second set of 3 covariates with same rule, but the correlation is 0.5, and the 3rd set of 3 covariates with same rule and correlation of 0.75. The final predictor is just a predictor with N(0,1), because I want to include an irrelevant predictor.

The results are:

|  |  |  |
| --- | --- | --- |
| Criterions: | Number of time the criterions matches the right model | Average MSE of test set. |
| R2 | 30 | 27.9 |
| R2adj | 18 | 28.3 |
| Cp | 41 | 27.9 |
| Press | 33 | 28.1 |
| AIC | 37 | 28.1 |
| BIC | 66 | 27.7 |

**Table 10**: simulation results of the mix of independent and dependent covariates setting

For this setting, the order of the model is still the same with setting 1, except AIC performs better compared to press, with higher number of time it matches the right model but the same average MSE of test set.

**Conclusion**

In all of 3 settings, BIC proves to outperform other criterions by a fair margin. Therefore, I believe BIC is the best criterion to find the right model. Cp is the second best criterion. AIC and press perform similarly. And R2adj performs the worst.

I also find that using an arbitrary stopping rule performs really well, especially when cases of correlated covariates, which indicates that human judgement should be used when choosing the right model