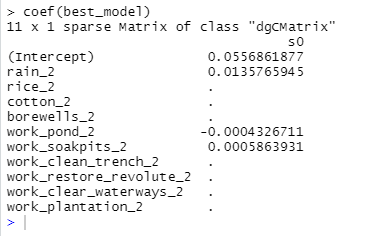
A lasso regression finds coefficients that minimize the sum of squared errors in the regression equation with an additional penalty term.

The penalty term results in the lasso regression shrinking the estimated regression coefficients towards zero and potentially setting coefficients on some variables exactly to zero, both of which help reduce over-fitting. The lasso, by setting some coefficients to zero, also performs variable selection. These shrinkage properties allow Lasso regression to be used even when the number of observations is small relative to the number of predictors.

However, directly using lasso regression can be problematic. Those lasso-estimated coefficients that are actually non-zero are typically underestimated, and lasso may mistakenly exclude variables with non-zero coefficients, particularly variables with moderate effects. Each of these phenomena generally causes significant regularization bias that adversely affects estimation and inference about coefficients. The omission of covariates with moderate but non-zero coefficients is especially problematic and results in omitted variable bias when these covariates are relevant predictors of the focal variable. In order to overcome such biases, we recommend using the “double-lasso” variable selection procedure which was explicitly designed to alleviate both sources of bias.

Two methods were used depending on whether or not there was an interaction between the independent variables by assumption. Lasso regression and double lasso regression. The results obtained by using Lasso regression are as follows (figure 1).

Select the best parameter prior to lasso regression. This is achieved through cross-validation. Select the lambda with the smallest mean two-squared error as the best penalty parameter. Best parameter is lambda= 0.02698796. As a result of our analysis, we get the following coefficient table:



The result of lasso method shows 3 variables (rain, pond and soakpits) affect water level. Here the affection of rain variable is the strongest positive. Then soakpits. Also pond affect negatively.

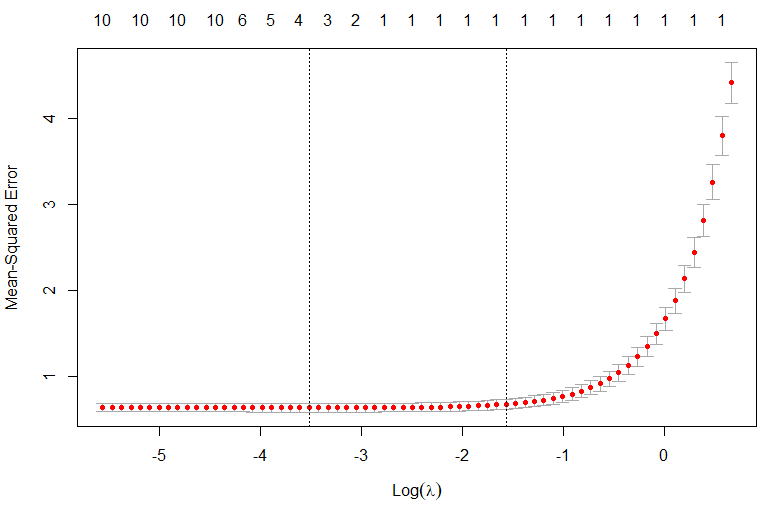
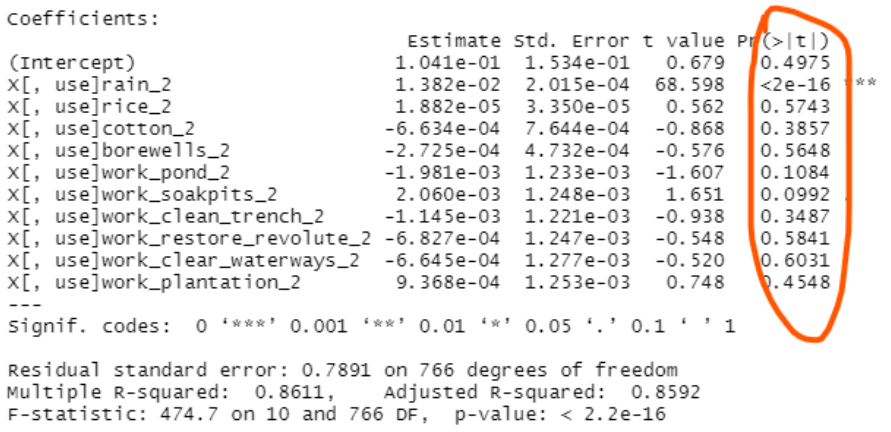


Figure1. MSE via Log()

Next, let’s see case that we suppose interaction. We used double lasso method at that time.



The figure above shows the coefficients estimated using double lasso regression, their standard deviations, t-statistics and p-value, and the hypothesis test results can also be seen.

Here most significant factor is rain variable. As a result of the hypothesis test, the second significant factor was soakpits, followed by the pond. All other variables have a significance level of 0.3 or higher, so these variables do not affect the water level.