Yuxuan Wang, December 2020, solution for final project Q3.

3 Variable selection

3.1 Implement three methods

3.1.1 Implement least square minimization

To implement least squares minimization:

$$\min_{\beta} \sum_{i=1}^{n} \left(Y_i - X_i^{\top}\hat{\beta}\right)^2$$

The derivation of the loss function is:

$$\frac{\partial \mathcal{L}}{\partial \beta} = -2X(Y_i - X_i^T\beta)$$

Using the derivation, we can calculate the gradient and iterate it until the optimal problem converges. The loss of least square minimization is showed in Figure 1.

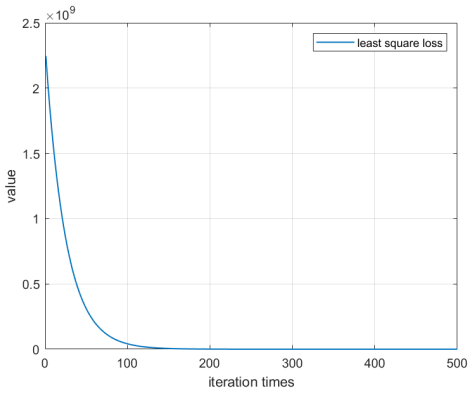
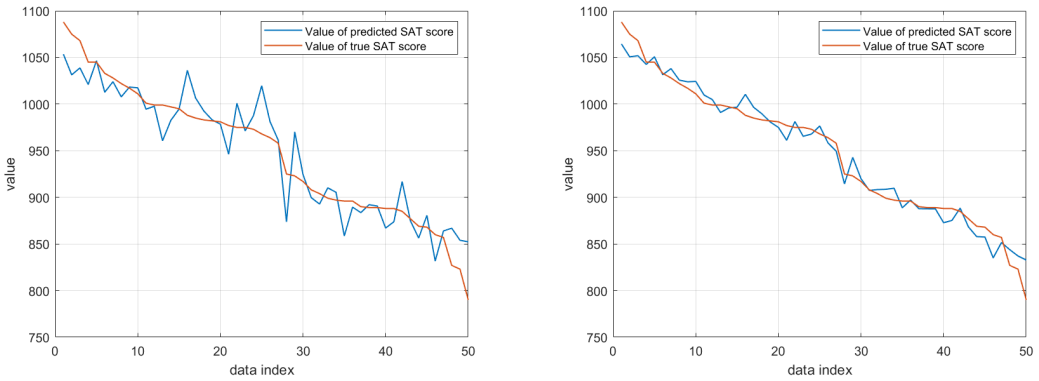


Figure 1: log loss of least square minimization

A problem that need me to pay attention is that we are given a dataset containing the variables of states, which is not a value. To tackle with this problem, there are three method:

- Discard the first column data.
- Code the first column data from 1 to 50 according to their index.
- Code the first column data randomly.

It is observed that method 3 did not show quite difference from method 1, so here we implement method 1 and 2. Figure 1 plots the prediction using least square minimization, on the training set we can observe that method 2 shows better performance. However, there might be overfitting, then we will use step-selection and cross validation to make a fair comparison. Here, we just implement the least square minimization.



(a) Prediction method 1

(b) Prediction method 2

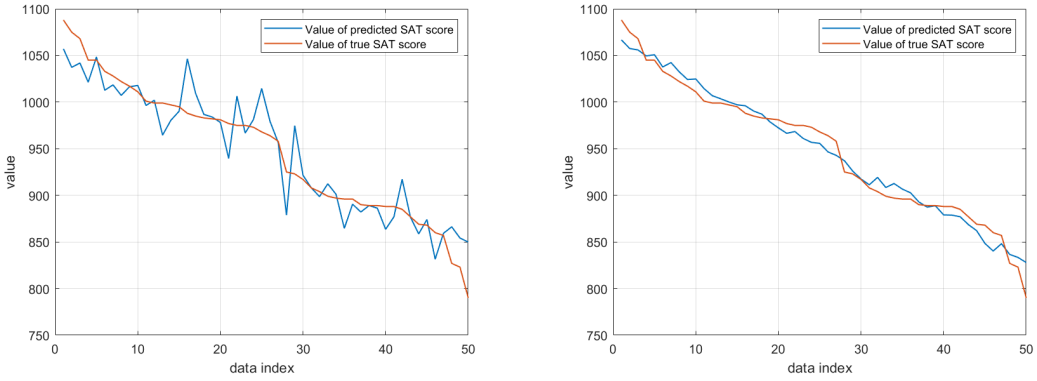
Figure 2: SAT prediction using least square minimization on training set.

3.1.2 Implement stepwise selection

For stepwise selection, it is started with two steps of forward selection and then alternates between one step of backward elimination and one step of forward selection.
From what we observed, the SAT scores decrease with the state index, which shows strong correlation with the coding method being adopted, so stepwize selection cannot exclude it, which may requires cross validation to check on it latter.
Variables selected by stepwise selections are shown as follows:

Table 1: Variables chosen by stepwise selection

Selected Variable(include states)	states	years	constant	
β	-4.6982	6.2570	966.3240	
Selected Variable(exnclude states)	years	expend	rank	constant
β	26.0952	1.8609	9.8258	-303.7243



(a) Prediction method 1

(b) Prediction method 2

Figure 3: SAT prediction using step minimization.

3.1.3 Implement lasso

Lasso is a method that accomplish regression and variable selection at a same time. To minimize:

$$\sum_{i=1}^{n}(Y_i - X_i^T\hat{\beta})^2 + \lambda||\beta||_1$$

Then the gradient is used to implement the iteration process:

$$\frac{\partial \mathcal{L}}{\partial \beta} = -2X(Y_i - X_i^T\beta) + \lambda \text{sign}(\beta)$$

For those covariates with small weights, we can regard it as unselected. For lasso, the relationship between covariates and the parameter λ is drawn in Figure.4(b). Variables with significant weights are presented in Table 2($\lambda = 5$).

Table 2: Variables chosen by lasso

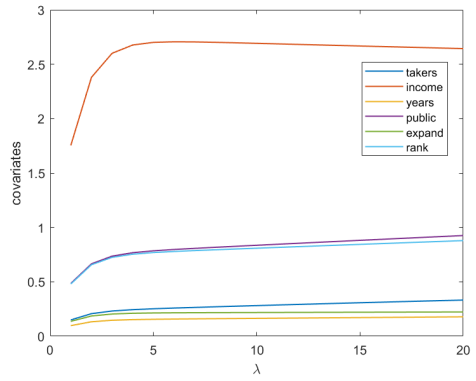
Selected Variable(include states)	states	takers	years	rank
Selected Variable(exnclude states)	takers	years	expand	rank

3.2 Comparison on three mothods

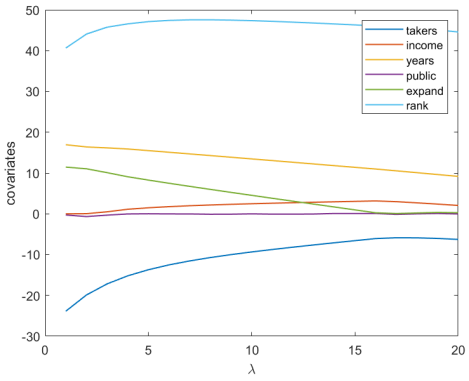
The error ϵ is calculated as follows:

$$\epsilon = \frac{1}{n_{\text{test}}}\sum_{i=1}^{n_{\text{test}}}\frac{|\text{True scores}_i - \text{Predicted scores}_i|}{\text{True scores}_i}$$

Using least square minimization, stepwise selection and lasso, the cross validation error are in Table.3. Least square minimization shows the lowest cross validation error and stepwise selection shows the highest because of fewer

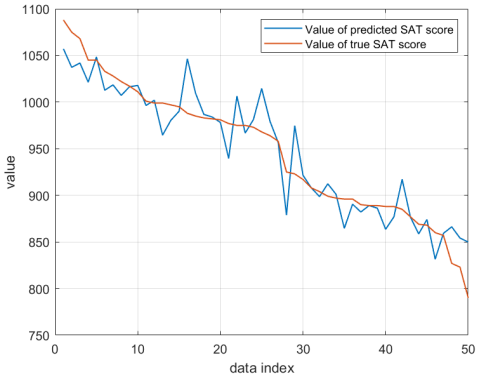


(a) Variable selected using lasso

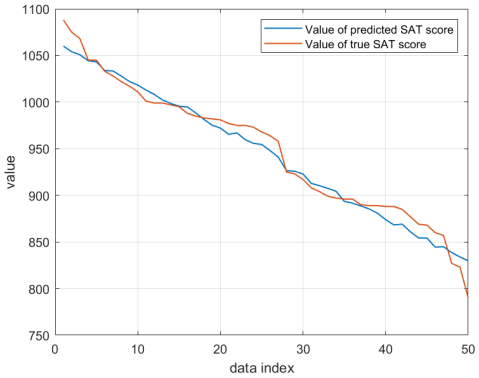


(b) Variable selected using lasso(With data normalization)

Figure 4: Implementing lasso



(a) Prediction method 1



(b) Prediction method 2

Figure 5: SAT prediction using lasso.

variables. It is worth noticing that coding states as a variable helps to reduce the cross validation error of both least square minimization and lasso, while stepwise selection would select less variables due to its significant influence and resulting in higher error rates.

Table 3: Cross validation error

Method	least square minimization	stepwise selection	lasso
Average cross validation error(coding states as a variable)	0.0137	0.0845	0.0147
Average cross validation error(excluding states)	0.0319	0.0379	0.0348