Model Averaging for Predicting the Exposure to Aflatoxin B$_1$
Using DNA Methylation in White Blood Cells of Infants

S Rahardiantoro$^1$, B Sartono$^1$ and A Kurnia$^1*$

$^1$Department of Statistics, Bogor Agricultural University

Email: anangk@apps.ipb.ac.id

Abstract. In recent years, DNA methylation has been the special issue to reveal the pattern of a lot of human diseases. Huge amount of data would be the inescapable phenomenon in this case. In addition, some researchers interesting to take some predictions based on these huge data, especially using regression analysis. The classical approach would be failed to take the task. Model averaging by Ando and Li [1] could be an alternative approach to face this problem. This research applied the model averaging to get the best prediction in high dimension of data. In the practice, the case study by Vargas et al [3], data of exposure to aflatoxin B$_1$ (AFB$_1$) and DNA methylation in white blood cells of infants in The Gambia, take the implementation of model averaging. The best ensemble model selected based on the minimum of MAPE, MAE, and MSE of predictions. The result is ensemble model by model averaging with number of predictors in model candidate is 15.

1. Background
In high dimensional data, which is popular in recent century, the classical regression analysis failed to bring the model of relationship two sets variables. This approach limited in dimension of data condition, number of observations has to higher than the number of predictors. Certainly, this condition is opposite with high dimensional data. Suppose the design matrix $X$ with dimension $n \times p$, high dimension condition occurs when $p \gg n$. Therefore, when facing with high dimensional regression cases, classical approach would be a bad choice.

Ando and Li, in 2014, proposed the alternative approach in regression analysis cases using high dimensional data that called model averaging (MA) [1]. Moreover, the MA approaches also applied in experimental design for supersaturated topics [2]. The main purpose of MA is to get the predictions of response variable accurately. MA takes some candidate model by the splitting process of predictor variables in the order of each marginal correlation with the response variable. The final predictions are from the averaging of the predictions of each candidate models.

Generally, high dimensional data condition occurs in the biological research, furthermore in the research of genetics. The one of popular cases are in DNA methylation that use the microarray technologies. An array can contain up to 20000 spots, which is each spot has been implanted with target DNA sequence. Because of that, it can’t be imagine how huge data used in this subject case.

This research focuses on the application of MA in the high dimensional data. The application taken based on the data of exposure to aflatoxin B$_1$ (AFB$_1$) and DNA methylation in white blood cells of infants in The Gambia by Vargas et al [3]. Aflatoxin is a contaminant of staple foods in sub-Saharan Africa, is a known human liver carcinogen and has been associated with stunting in infants. MA
applied to get the best prediction of AFB, from DNA methylation based on MAPE, MAE, and MSE criteria. Number of predictors in the model candidates \( (m) \) has been set in some options, therefore the minimum of MAPE, MAE and MSE values indicates the predictions of MA with specify \( m \) to be the best choice. In practice, this research using R software in implementing all of process.

2. Methodology
In this section would be described about the data conditions, algorithm process of MA, and also the value of best-model criterias (MAPE, MAE, and MSE) [4].

2.1. Data
The data of this research is from the case based on Vargas et al [3] in the previous section. There were 124 pregnant mothers to be subject of research. Variable of AFB, took from those pregnant mothers in first trimester. Their 3\textsuperscript{rd} – 6\textsuperscript{th} month infants have been studied to take the DNA methylation in their white blood cell that contained 485577 profiles. Based of that, the response variable, AFB, has the number of observations \( n = 124 \), and the predictor variables, DNA methylation, have the number of profiles \( p = 485577 \), which is include in the high dimensional data (\( p \gg n \)).

The first step before the analysis process is data cleaning. This process was taken to avoid the missing values in the analysis, that is guarantee the MA process well done. The dimension of data after the cleaning process is \( n = 118 \) and \( p = 480050 \). In the practices, these data divide into data training (for modeling process) and data testing (for prediction process). Therefore, MA approach applied in these data.

2.2. Model Averaging Methods
One of necessary point in the MA is determining the number of predictors in the candidate model \((m)\). In this research take 8 numbers, which is \( m = 5,10,\ldots,35,40 \). The averaging process taken with the specify weight \( w_i \), AIC weight which is based on the value of AIC in each model candidates. Suppose there are \( k \) model candidates, therefore the \( i \) – th AIC weight follows

\[
w_i = \frac{\exp \left( \frac{1}{2} a_i \right)}{\sum_{i=1}^{k} \exp \left( \frac{1}{2} a_i \right)}
\]

where \( a_i \) denotes the value of AIC in the \( i \) – th model candidates, and \( w_i \geq 0; \sum_{i=1}^{k} w_i = 1 \) [3].

The ensamble predictions of MA define by the weighted averaging of each model candidate predictions. Suppose \( \mathbf{y}_i \) is the prediction vector of \( i \) – th candidate model, therefore

\[
\mathbf{y} = \sum_{i=1}^{k} w_i \mathbf{y}_i
\]

become the ensamble prediction vector.

Furthermore, in the model candidate preparation, the selection of predictor variables is the essential process in MA. Based on the proposed MA by Ando and Li (2014), predictors selection in model candidate is done based on the rank of marginal correlation between all of predictors and the response variable [1].

2.3. Evaluation Criteria
The evaluation criteria in this research are based on the value of Mean Absolute Percentage of Error (MAPE), Mean Absolute of Error (MAE), and Mean Square Error (MSE). These criteria values selected because include in the most popular prediction quality measurements that useful in regression cases [5]. The minimum values of MAPE, MAE, and MSE indicate that the model selected give the best predictions. The formulas of MAPE, MAE, and MSE as follow
MAPE = \left( \frac{\sum_{t=1}^{n}|y_t - \hat{y}_t|}{n} \right) \times 100\% \quad (3)

MAE = \frac{\sum_{t=1}^{n}|y_t - \hat{y}_t|}{n} \quad (4)

MSE = \frac{\sum_{t=1}^{n}(y_t - \hat{y}_t)^2}{n} \quad (5)

where \( t \) indicates the index of number of observations, \( t = 1, 2, ..., n \).

3. Result and Discussion

In \( n = 118 \) and \( p = 480050 \) data, and \( m = 5, 10, ..., 35, 40 \), table 1 shows the result of measurement accuracy of prediction.

| \( m \) | MAPE | MAE | MSE |
|-------|------|------|-----|
|       | Data training | Data testing | Data training | Data testing | Data training | Data testing |
| 5     | 2.02 | 0.79 | 36.76 | 30.67 | 2289.87 | 2723.57 |
| 10    | 1.49 | 0.94 | 30.47 | 40.43 | 1560.06 | 4125.55 |
| 15    | 1.47 | 0.78 | 28.29 | 40.49 | 1166.02 | 8203.09 |
| 20    | 1.48 | 1.43 | 24.37 | 46.91 | 1036.51 | 3117.30 |
| 25    | 1.50 | 1.59 | 25.33 | 52.94 | 1023.13 | 4451.83 |
| 30    | 0.89 | 1.36 | 17.70 | 48.47 | 572.00 | 5570.61 |
| 35    | 1.09 | 1.72 | 18.02 | 72.43 | 549.17 | 17214.61 |
| 40    | 0.91 | 2.48 | 16.36 | 78.43 | 440.24 | 15721.91 |

*the yellow highlight indicate the lowest values

The best accuracy was determined based on the lowest MAPE, MAE and MSE in the data testing. Based on those results, \( m = 15 \) has the lowest MAPE value, and \( m = 5 \) has the lowest MAE and MSE. Figure 1 shows the \( m = 15 \) and \( m = 5 \) prediction with the actual of AFB1 variable.

**Figure 1.** Scatter plot of the \( m = 15 \) and \( m = 5 \) prediction with the actual of AFB1 variable
The charts on Figure 1 suggest the following general pattern. Model with $m = 15$ has better pattern than model with $m = 5$, it also supported by the correlations between the actual values and the prediction values 0.523. Based on this research, we suggest the model averaging as an alternative when the data has high dimension, in the prediction AFB$_1$ using DNA methylation of white blood cell, we can choose the $m = 15$ model averaging.

4. Conclusion
The MA can be an alternative approach in the case of high dimensional regression. The prediction of AFB$_1$ from pregnant mothers using the DNA methylation on white blood cells of their infants can be solved using MA. The best model based on the minimum value of MAPE, MAE, and MSE, also in addition the highest correlation between the prediction values and the actual values, is MA with $m = 15$.

References
[1] Ando T and Li K C 2014 *Journal of the American Statistical Association* **194** 254-265.
[2] Salaki D T, Kurnia A and Sartono B 2016 *IOP Conf. Series: Earth and Environmental Science*. **31** 1-7.
[3] Vargas H H, Castelino J, Silver M J, Salas P D, Cros M P, Durand G, Kelm F L C, Prentice A M, Wild C P, Moore S E et al. 2015 *International Journal of Epidemiology* 1-11.
[4] Myttenaere A D, Golden B, Grand B L, Rossi F 2015 Using the Mean Absolute Percentage Error for Regression Models. *Proceedings of the 23-th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN 2015).* hal-01162980.
[5] Claeskens G, Hjort N L 2008 *Model Selection and Model Averaging* (New York (US): Cambridge University Press)