Predicting Response Time of Questions Using Linear Mixed-effects Model

Luyao Peng
Awada Tech LLC.
luyaopeng.cn@gmail.com

ABSTRACT

Estimating response time (RT) to questions in exam or homework practices is becoming more important in online learning platforms to improve the learning experience and item allocation. To predict RT, we applied the Empirical Best Linear Unbiased Predictor (EBLUP) technique from the linear mixed-effects model to predict RT to each question. The proposed method provides more precise response time predictions in terms of mean absolute errors, correlation coefficient, and close-to-1 ratio counts compared with feature-based linear regression and 95 quantile predictions.

Keywords

Linear Mixed-effects Model, EBLUP, Response Time Prediction, Bayes Estimator

1. INTRODUCTION

As digital communication devices have become ubiquitous, online testing and internet-based learning platforms are becoming more popular testing and learning forms.

In those applications, predicting the response time (RT) needed to complete each question in learning practices and exams has implications for both reasonable item allocations in exams and improving learning experiences and efficiencies. In learning practices and adaptive exams, it is often desirable to allocate items across a range of RTs, because allocating too little time makes the test and practices speeded and affect learning experiences whereas allocating too much time makes it inefficient.

There are three commonly used models for RT predictions of test items: Natural Language Processing (NLP) techniques based on text features; feature-based machine learning regressions such as Random Forest and Neural Networks (RT as the sole dependent variable); lognormal model based on RT data and hierarchical RT model based on both response matrix and RT data (RT and response accuracy as dependent variables).

However, NLP methods require high-dimensional word embeddings and linguistic characteristic feature engineering; machine learning methods also require complicated feature engineering to predict RT; lognormal and hierarchical RT models require larger sample sizes since they strictly assume the distributions of RT and the Item Response Theory (IRT) model of response data, moreover, RT models not only estimate item speed parameters, they also focus on estimating item difficulty and student ability parameters.

Studies in predicting RT of questions have still been underrepresented in the test and learning activity constructions, especially when predicting RT to a new item with few historical RT logs. Therefore, in the current study, we apply the Empirical Best Linear Unbiased Predictor (EBLUP) in the mixed-effects model to predict RT to items even when their RT samples are small.

In summary, the contribution of the work is three-fold: (1) the first study in applying EBLUP to item RT predictions; (2) predicting RT when the sample size for the item is small; (3) demonstrating the precision of EBLUP comparing with other RT models and 95 quantile methods in online learning programs using Riiid data for full and small samples, respectively.

2. BACKGROUND WORK

In this section, we will briefly review common methods in RT prediction and the mixed-effects model that will be applied in RT prediction.

2.1 Common Methods in RT Prediction

There are 3 types of common methods in RT prediction, they are Natural Language Processing (NLP) technique, Machine Learning regression and Response Time Lognormal models.

NLP technique analyzes linguistic features [2] and word embeddings [7] generated from item text to predict RT. The drawbacks of the NLP methods in RT prediction are that they require high-dimensional feature engineering and large sample sizes to achieve desirable performances.

Researchers from the statistical measurement realm estimate RT by modeling it as a set of latent parameters in the Response Time Lognormal models, the log-response time is then regressed to the RT, item difficulty and student abil-

L. Peng. Predicting response time of questions using linear mixed-effects model. In B. Paaßen and C. D. Epp, editors, *Proceedings of the 17th International Conference on Educational Data Mining*, pages 877–881, Atlanta, Georgia, USA, July 2024. International Educational Data Mining Society.

https://doi.org/10.5281/zenodo.12729982

ity parameters. However, The purpose of the lognormal RT model is not only estimating the RT for each item. Instead, it investigates the ability of a person on a set of test items and improves model fit to RT data. In addition, RTL models are time consuming due to the large amount of parameters to be estimated.

Previous studies also used Machine Learning methods to predict RT. For example, [3] extracted 10 features from the student's interaction with the mouse in the test and trained Random Forests, Neural Networks, Linear Regression, and Gaussian Process regression to predict the completion time of each question. Random Forests and Neural Networks give the smallest average prediction errors in predicting the completion time compared to the other two methods.

In this study, We only focus on the time-efficient methods, therefore, we will compare our proposed method with Machine Learning methods such as the Linear regression and Gradient Boosting regression.

2.2 Linear Mixed-Effects Model and EBLUP Predictions

2.2.1 The Linear Mixed-Effects Model

Suppose J observations and p independent features, each observation is nested within one of the I groups, then the LMM is defined as:

$$y_{ij} = \mu + X_i\beta + b_i + e_{ij}, \tag{1}$$

where y_{ij} is an observation j in group i, μ is the expected value for the dependent variable, X_i and β are the $1 \times p$ independent features vector for group i and the corresponding coefficient vector of dimension $p \times 1$, respectively; b_i is the random effects for group i and is assumed to follow iid $N(0, \sigma_b^2)$, e_{ij} is the random error for observation j and is assumed to follow iid $N(0, \sigma_e^2)$. The model in Equation 1 is also called the nested-error mixed-effects regression model [6].

The LMM defined in Equation 1 can also be rewritten as:

$$\begin{gathered} y_{ij} \overset{iid}{\sim} N(y_i, \sigma_e^2) \\ \text{where } y_i \overset{iid}{\sim} N(\mu + X_i\beta, \sigma_b^2) \end{gathered} \tag{2}$$

which is a typical form of a hierarchical Bayesian model under normal distributions.

2.2.2 The BLUP and the EBLUP of LMM

Since we are interested in predicting y_i in model 2, the true value of the dependent variable for group i, then, the Best Linear Unbiased Predictor (BLUP) can be derived as the following based on [6]:

$$\hat{y}_i^{BLUP}(\sigma_e^2, \sigma_b^2) = (1 - \gamma_i)\left[\hat{\mu} + \bar{X}_i\hat{\beta}\right] + \gamma_i\bar{y}_i. \tag{3}$$

where $\bar{X}_i$ is the sample mean of the independent feature for group i; $\bar{y}_i$ is the sample mean of the dependent variable for group i; $\gamma_i = \sigma_b^2\left(\sigma_b^2 + \sigma_e^2 n_i^{-1}\right)^{-1}$; $\hat{\mu}$ and $\hat{\beta}$ are the ordinary least square (OLS) estimators of the coefficient parameters in model (1).

It is interesting to note that γ_i is a weighting parameter between the least square estimator $\hat{\mu} + \bar{X}_i\hat{\beta}$ and the sample mean estimator $\bar{y}_i$ for the true value of y_i. If $\sigma_e^2 n_i^{-1}$ is large (either large random errors in the observed y_{ij}'s or small sample size in group i), γ_i will become small, therefore less weight will be given to the sample mean predictor $\bar{y}_i$ and more weight will be given to the OLS estimator $\left[\hat{\mu} + \bar{X}_i\hat{\beta}\right]$; on the other hand, if σ_b^2 is large or sample sizes are sufficient, γ_i will be relatively large, and more weight will be given to the group-specific sample mean $\bar{y}_i$.

BLUP is used directly when the variance components σ_e^2 and σ_b^2 in (3) are known, but they are unknown for real data set and need to be estimated using Restricted Maximum Likelihood (REML) method [5] under the model in Equation (1), one obtains what is referred to as the Empirical Best Linear Unbiased Predictor (EBLUP) [4] of t_i as

$$\hat{y}_i^{EBLUP}(\hat{\sigma}_e^2, \hat{\sigma}_b^2) = (1 - \hat{\gamma}_i)\left[\hat{\mu} + \bar{X}_i\hat{\beta}\right] + \hat{\gamma}_i\bar{y}_i. \tag{4}$$

The formulation of BLUP/EBLUP in LMM is beneficial in predicting the true value of interested features at the group level especially when the sample sizes are small in that group. In RT prediction, there are many new items that don't have sufficient RT history but still need RT predictions for practical implementations. In this case, the BLUP/EBLUP is a suitable method for solving the small data issues for new item RT prediction, this is also the motivation of our study to apply LMM in RT prediction.

3. APPLYING EBLUP OF LMM IN RT PREDICTION

3.1 The LMM Model in RT Prediction

To apply the theory of LMMs to predict RT to test items, we replace y_{ij} in (2) by t_{ij}, the log RT of student j to item i :

$$t_{ij} = \mu + X_i\beta + b_i + e_{ij}, i = 1, \ldots, I, j = 1, \ldots, J \tag{5}$$

where μ is the expected log RT for the whole item population, X_i is the independent feature of each item, it represents the students' accuracy rate for item i, its coefficient is β; b_i is the random effect for item i, indicating the variation of RT for item i from the expected RT μ, it is assumed that $b_i \overset{iid}{\sim} N(0, \sigma_b^2)$; e_{ij} is the random errors for t_{ij}, it is assumed that $e_{ij} \overset{iid}{\sim} N(0, \sigma_e^2)$. Note that more features can be included in the LMM; here, we only include the accuracy rate for each item. Hence, there is only one β coefficient in the model.

3.2 The EBLUP Predictors

Given the formulation of EBLUP in Equation (4), we can obtain the EBLUP for predicting the true RT value of each item by the following:

$$\hat{t}_i^{EBLUP}(\hat{\sigma}_e^2, \hat{\sigma}_b^2) = (1 - \hat{\gamma}_i)\left[\hat{\mu} + \bar{x}_i\hat{\beta}\right] + \hat{\gamma}_i\bar{t}_i. \tag{6}$$

$\hat{t}_i^{EBLUP}$ is the predictor for the expected log RT for item i, that is the expected time by which students can finish completing an item, we still need to obtain the RT by which the majority of the students can finish the item, therefore, we need to predict the RT that 95% of the students can finish item i in this study.

3.3 Upper Bound of RT by EBLUP

Using $\hat{t}_i^{EBLUP}$ in Equation (6), the final RT prediction is:

$$\hat{t}_i^{EBLUP'} = \hat{t}_i^{EBLUP} + z * \hat{\sigma}_{RTi}, \tag{7}$$

where $\hat{\sigma}_{RTi}$ is the sample standard deviation (SD) of log RT for item i, from the historical RT data; z is how many SD away from the predicted expectation of RT under the sample distribution of the observed log RT's for item i. We determine the value of z via cross-validation procedure, which will be discussed in Section 4.3.

4. REAL DATA AND RESULTS

We demonstrated the RT prediction performances of EBLUP and compared the predictor with other predictors: Weighted Least Square (WLS) predictor, 95 quantile predictor, and gradient boosting regressor (GBR) predictor using the Riiid dataset [1] [1]. Because we don't have the question's texts and we only focus on time-efficient methods in this study, we exclude the NLP and RT models in our real data comparison.

4.1 Riiid Data and Data Preprocessing

Riiid data is students' test answer data from a complete education app, the data include student's historic performance, the performance of other students on the same question, metadata about the question itself, and it contains test answering data about 1M+ students in time-series format.

We preprocessed the Riiid data by removing the records without responses, removing the records for the lecture content_type ('content_id=1'), keeping the first record for duplicated rows if the same user_id answers a 'content_id' more than once; we also removed the records if the RT is too fast (RT < quantile 0.01) or too slow (RT > quantile 0.99), presumably because too fast RT indicates guessing or cheating, and too slow RT indicates learner being offline or distracted; since we use the historical RT data to compute predictors and test the precision of the predictors on the test data, we only keep the items with RT records greater than 20. After preprocessing, the data contains 368420 users and 4712 items.

Since the data is time-series data within each item, we can't use random sample cross-validation to determine the optimal values of the parameters in each method. In this case, we sorted the RT (in milliseconds) for each item by 'timestamp', and then keep the first half of the records within each item as the training data (4712 items and 189707 users), use the first half of the remaining data as validation data (4712 items, 95208 users) and the remaining data within each item as the test data (4712 items, 95272 users). The data contains variable 'prior_question_elapsed_time', so we shifted the variable by -1 and use the shifted 'prior_question_elapsed_time' as the RT of each student for each item. In addition, we compute a new variable 'accuracy_rate' for each item using the variable 'answered_correctly' / the number of responses for each item, and use the accuracy rate as the independent variable in all feature-based methods. For WLS and GBR, in addition to 'accuracy_rate', we also add question-specific independent variables 'bundle_id' and 'part' and convert them to categorical variables.

4.2 Evaluation Criteria

We compared the performance of the EBLUP with the other three methods by using three evaluation metrics:

1. Mean Absolute Errors (MAE): which is the average of the absolute difference between the predicted RT and the observed RT across all items in test data.

2. Close-to-1 Ratio Counts (RC): we allow a certain degree of discrepancy of the prediction, hence, we compute the ratio between the prediction and the observed 95th RT in the test data (being closer to 1 indicates that the prediction and the observed 95th RT are the same, the closer to 1, the more precise the prediction).

Then, we group the ratios to 7 categories: $[0, 0.5]$, $[0.5, 0.8]$, $[0.8, 1.0]$, $[1.0, 1.2]$, $[1.2, 1.5]$, $[1.5, 2]$, $[> 2]$ and count the frequencies within each ratio interval, we expect more ratios within the intervals of $[0.8, 1.0]$ and $[1.0, 1.2]$.

3. Correlation Coefficient (CC): we also compare the correlations between the predictions and the observed 95th RT for each method.

Because EBLUP is especially beneficial when the sample sizes of the item is small (such as new items), therefore, we compute the three metrics of each method for the full item data and the small item data, respectively. The full item data contains all the items, while the small item data are the items with answer frequencies < 50.

4.3 Methods Implementation

4.3.1 EBLUP

In the mixed-effects model of EBLUP, we regress t_{ij}, the log of the shifted 'prior_question_elapsed_time' to the accuracy rate for each item X_i, we use 'question_id' as the unit of the random effects b_i. We use the resulting $EBLUP$ computed by Equation (6) as the predicted expectation of the RT for item i.

To find the upper bound of RT for each item, we tune the values of z from 1.5 to 2 by step of 0.01 to find out the best z in Equation (7) in giving the most close-to-1 RCs in the validation data, results show $z = 1.78$ predict the RT on validation data the best.

4.3.2 WLS, 95 Quantile and GBR

We compute the 95th RT quantile for each item and use the log-transformed quantile as the direct predictor for the RTs of the items on test data; We use the 95th quantile RT for each item as the dependent variable, 'accuracy_rate',

[1] https://www.kaggle.com/competitions/riiid-test-answer-prediction/data

'bundle_id', 'part' as the independent variables in WLS and GBR. We obtained the best parameters through a 1-fold cross-validation for WLS and GBR predictors based on the sorted timestamp. Check our code for detailed implementations [2].

4.4 Results

We report the 3 evaluation metrics for the full item data and the small item data (with <50 RT records), respectively.

Table (1) and (2) show that EBLUP produces the smallest MAE on full data, and is slightly smaller MAE compared to GBR for small item data. EBLUP also produces the smallest Standard Deviation (SD) of MAE for small item data. Table (3) shows the CCs for the 4 methods, EBLUP has the highest correlation with the observed 95th RT on both full data and small item data. Figure (1) shows that WLS and GBR tend to underestimate the long RT items.

Table 1: MAE Comparisons for All Items (in milliseconds)

	Quant 95	EBLUP	WLS	GBR
MAE	6572	6046	6937	6482
SD of MAE	6392	5755	6236	5723

Table 2: MAE Comparisons for Small Items (in milliseconds)

	Quant 95	EBLUP	WLS	GBR
MAE	7063	6417	6925	6570
SD of MAE	6657	6022	6439	6392

Table 3: Correlation Comparisons

	Quant 95	EBLUP	WLS	GBR
All items	0.678	0.702	0.678	0.676
Small items	0.644	0.669	0.650	0.600

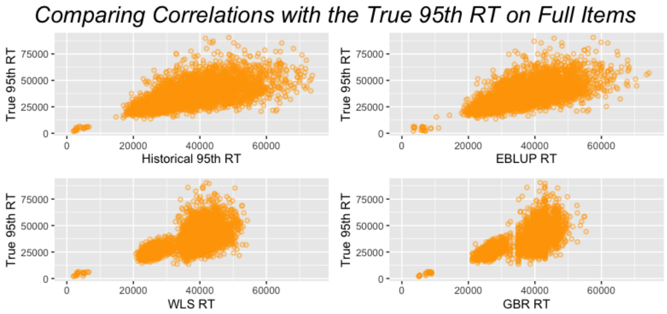

Figure 1: Scatterplots for 4 methods with CC on Full Item Data

We compare the ratio counts within each of the 7 ratio intervals for the full data and the small item data, respectively. EBLUP has the most ratios (3202 and 2512) within the $[0.8, 1]$ and $[1, 1.2]$ intervals for the two data sets; GBR also performs well on both data.

Figure (2) shows the density comparisons of the ratios from the 4 methods, the dotted vertical lines represent the 0.8 and 1.2 boundaries. For the full data, EBLUP has obviously higher density between [0.8, 1.2], and slightly higher density on the small item data compared to GBR. WLS has the lowest ratio density on both data.

Table 4: Comparing Counts on Full Data

	Quant 95	EBLUP	WLS	GBR
[0,0.5]	9	9	11	8
[0.5,0.8]	527	547	657	401
[0.8,1]	1511	1674	1495	1469
[1,1.2]	1591	1528	1357	1725
[1.2,1.5]	818	781	944	893
[1.5,2]	238	156	205	203
>2	18	17	43	13

Table 5: Comparing Counts on Small Data

	Quant 95	EBLUP	WLS	GBR
[0,0.5]	9	9	9	6
[0.5,0.8]	489	490	544	378
[0.8,1]	1187	1231	1149	1156
[1,1.2]	1231	1281	1147	1322
[1.2,1.5]	750	741	839	799
[1.5,2]	230	145	193	221
>2	18	17	33	32

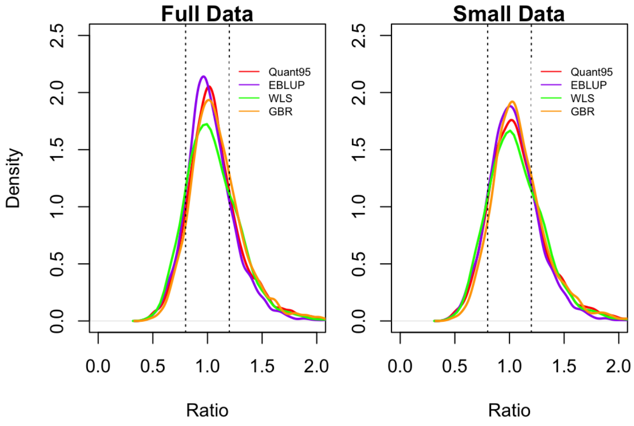

Figure 2: Ratio Densities on Full and Small Data

We see that EBLUP tends to have more prediction ratios in [0.8, 1], because when the sample is large for each item, EBLUP gives more weight on the sample estimate of the RT, it is more similar to Quant 95; when the sample is small, such as new items, EBLUP gives more weight to the generalized least square estimate of the whole item population.

Recall that EBLUP only uses 'accuracy_rate' as the independent variable, while WLS and GBR include other independent variables. It is suggested to use EBLUP when the independent features are limited or difficult to extract. On the other hand, if more item features are included in the LMM, WLS and GBR models, the prediction accuracy will be improved.

[2] https://github.com/pengluyaoyao/Predicting-Response-Time-of-Questions-Using-Linear-Mixed-effects-Model/tree/main

5. CONCLUSION

Our work applied EBLUP in the linear mixed-effects model to predict RT for each item. EBLUP is a predictor by weighting between the sample estimate and the generalized least square estimate of RT for each item, it is the empirical best linear unbiased predictor for item-level RT predictions even when the sample size is small.

Results show that EBLUP outperforms or slightly outperforms the other three methods on both full item data and small item data in terms of MAEs, CCs and RCs. EBLUP has better performance on the full item data compared to the small item data. Due to the formulation of EBLUP in Equation (4), it tends to have shorter RT predictions compared to other methods.

Although EBLUP performs relatively well in different metrics, this study has limitations: 1. we didn't include categorical features in the LMM, however, categorical features can be added to improve the prediction performance; 2. unlike WLS and GBR which uses 95th quantile RT for each item as dependent variable directly, EBLUP can only predict the expectations of the interested variables, it cannot predict 95th RT quantile directly, therefore, we have to use Equation (7) to compute the upper bound prediction of RT for each item. 3. We only apply one-fold cross-validation based on the sorted timestamp, we can apply the cross-validation for time-series data in future study to find the optimal values for the parameters in GBR and EBLUP.

6. ACKNOWLEDGMENTS

This study is sponsored by Shanghai Pujiang Program, through Grant No. 22PJ1421800.

7. REFERENCES

[1] e. a. Addison Howard. Riiid answer correctness prediction, 2020.

[2] P. Baldwin, V. Yaneva, J. Mee, B. E. Clauser, and L. A. Ha. Using natural language processing to predict item response times and improve test construction. *Journal of Educational Measurement*, 58(1):4–30, 2021.

[3] D. Carneiro, P. Novais, D. Durães, J. M. Pego, and N. Sousa. Predicting completion time in high-stakes exams. *Future Generation Computer Systems*, 92:549–559, 2019.

[4] J. Jiang. Asymptotic properties of the empirical blup and blue in mixed linear models. *Statistica Sinica*, pages 861–885, 1998.

[5] C. E. McCulloch and S. R. Searle. *Generalized, linear, and mixed models*. John Wiley & Sons, 2004.

[6] N. N. Prasad and J. N. Rao. The estimation of the mean squared error of small-area estimators. *Journal of the American statistical association*, 85(409):163–171, 1990.

[7] K. Xue, V. Yaneva, C. Runyon, and P. Baldwin. Predicting the difficulty and response time of multiple choice questions using transfer learning. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 193–197, 2020.