Mutilevel Analysis on the Battleship Numberline Study Part II

Zuojun Gong

December 16, 2016

Thinking about the response variable reacTime

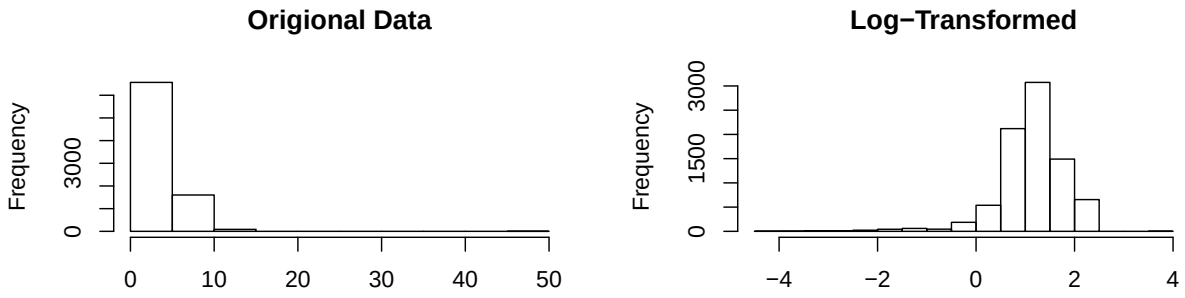


Figure 1: Distribution of the ReacTime varaiable: Origional vs log-Transformed

When comparing the distribution between the original reaction time variable and the log-transformed variable, we could see that the original data was skewed to the right with several large leverage points. Since we were building a standard hierarchical model with normal errors and normal random effects, it was reasonable to have our response variable to be normally distributed.

As we take a closer look at the consistency between reacTime and HitType, we found out that the time limit was not very adhered to in the data. The inconsistency was in timing cutoff. Of all the cases that have reacTime being recorded under 10 seconds, which was under the time limit, 164 of entries were marked as "Time Out!!" despite having reac time under 10 seconds. When we examine the reactime, it appears that most cases had the reac time around 9.9 seconds, with several other cases to as low as 8.6 seconds. However, the reverse case where reaction time was greater than 10 seconds and not marked as time out does not exist. It was possible that this was caused by communication or process delay, but the real reason unknown.

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
8.657	9.953	9.984	9.931	9.997	9.999

Given the game mechanism, we know once the game was timed out, the player was unable to answer the question. This will create a censor effect in our data, where we do not know the exact time it takes for the player to respond, but only knowing that the person took more than 10 seconds. The cutoff in our data could be seen in the log-transformed histogram of reac time in figure 1. Our residual plots will appear to have a sharp cutoff around log(10) and it will not appear as random as we expect them to. Given this situation, it might affect our random effects model because it "draws" the players whose response times were closer to each other by enforcing the same cutoffs, epically for players who takes longer (closer to 10 seconds) to respond on average.

We decided to make all Timed Out instances react time to 10 seconds in order to keep the to formalize these

cases. It would be unfair to exclude them from the dataset, since timed out usually suggests that the person was unable to provide a solution within 10 seconds, and this was equivalent to providing an incorrect answer. Thus, by making all timed out cases react time to 10 seconds, we corrected the issue in cases where timeout occurred right before 10 seconds, and cases where reactime was still recorded after the timeout. Also, it was helpful to know that all timed out cases in our dataset were marked as incorrect, which was consistent with the game design.

Ordinary linear regression for question effects

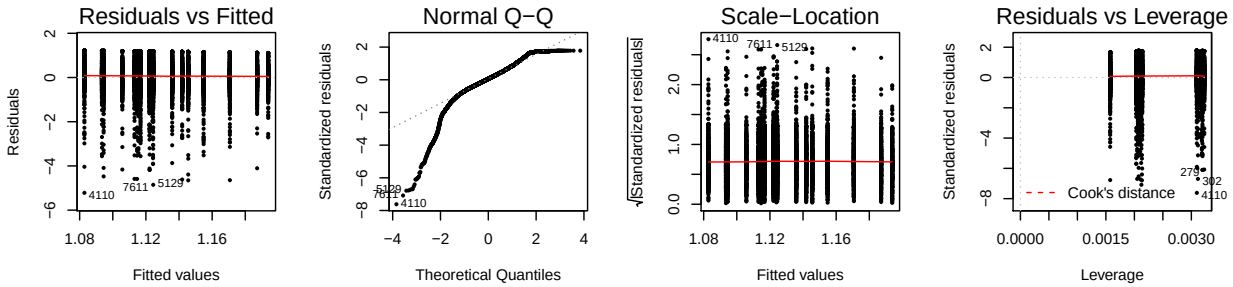


Figure 2: Diagnostics for Linear Model of lrt ~ CurreutQuestion

First, we fit a linear model on the log reaction time by using the type of current question as factors. From our linear model's diagnostics plots, we could see that this model mostly satisfies zero expectation, residual independence and constant variance assumption of the residuals. As expected, we see a clear cutoff on the upper end of the residual plots caused by the timing cutoff. We could see from the normal QQ plot that the residuals does not fit the normality assumption well, and it has tails on both end of the curve. In terms of model parameters, each factor corresponds to a question, and its parameter was the log of expected time to answer that specific question. In our example, the unbiased estimate was given by the average log-reaction time of each question. To put the parameter into perspective, for example, our model suggests that the expected log response time for question 0.1 was 1.1455. By taking the intercept out of the model, it was easier to compare the coefficients across all current questions instead of having to compute the relative values by using the reference intercept.

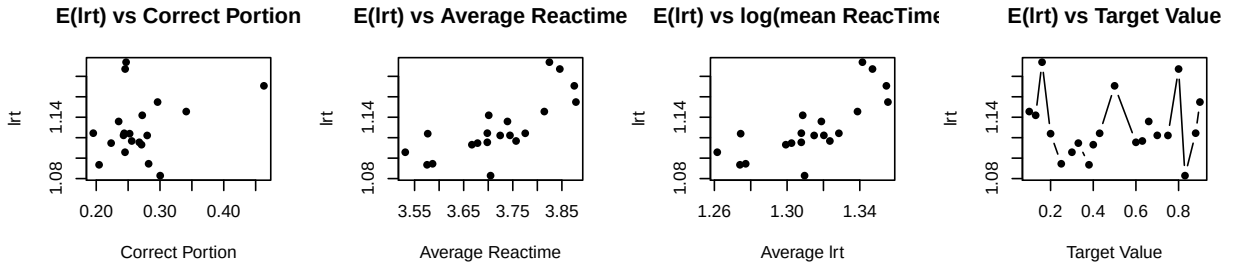


Figure 3: Expected Log-Reaction Time vs Question Easiness, average reaction time, log of the average reaction time, and the target value of the questions

In order to visualize the relationship between coefficients of the model and other variables, we created scatter plots of coefficient against the questions' easiness (portion of players who got the question right), mean of all reaction time and log reaction time and the numeric value of the question. It was not obvious whether there existed a relationship between the expected log reaction time vs corrected portion and target value.

The expected log react time vs Average React time was very similar to the plot where we took the log of the average reaction time. Also, it was interesting to see that the expected log reaction time did not appear to have an obvious relationship with the average reaction time, this was possibly caused by the difference in sample sizes in each question. This also indicated that the average of log response time was not the same as the log of average response time.

We preformed variable selection with stepwise AIC with both forward and backwards search. The smallest model that we considered was previous model with log reaction time as the response variable and current question as predictor variable. The largest model we have considered used log reaction time as the response variable, and current question, whether guide was enabled, the current level number, the average accuracy of the subject at the time of answer, the total star count, the current accuracy of the subject, number of items played, and the best time as predictor variables.

	AOR	CI lb	CI ub %	Pval
(Intercept)	1.612	1.488	1.746	<0.001
factor(resp)1	1.13	1.095	1.166	<0.001
factor(currentQuestion)0.13	1.035	0.962	1.112	0.357
factor(currentQuestion)0.16	1.116	1.029	1.211	0.008
factor(currentQuestion)0.2	1.065	0.982	1.155	0.131
factor(currentQuestion)0.25	1.041	0.969	1.119	0.274
factor(currentQuestion)0.3	1.064	0.981	1.155	0.136
factor(currentQuestion)0.33	1.038	0.966	1.116	0.311
factor(currentQuestion)0.38	1.074	0.989	1.165	0.09
factor(currentQuestion)0.4	1.052	0.97	1.141	0.22
factor(currentQuestion)0.43	1.051	0.978	1.129	0.178
factor(currentQuestion)0.5	1.026	0.959	1.098	0.457
factor(currentQuestion)0.6	1.069	0.985	1.159	0.111
factor(currentQuestion)0.63	1.044	0.971	1.123	0.24
factor(currentQuestion)0.66	1.083	1.007	1.165	0.033
factor(currentQuestion)0.7	1.049	0.967	1.138	0.248
factor(currentQuestion)0.75	1.043	0.969	1.121	0.263
factor(currentQuestion)0.8	1.1	1.015	1.193	0.021
factor(currentQuestion)0.83	0.996	0.919	1.08	0.921
factor(currentQuestion)0.88	1.043	0.97	1.122	0.256
factor(currentQuestion)0.9	1.031	0.959	1.109	0.402
factor(isGuidesEnabled)TRUE	1.049	1.018	1.08	0.002
factor(currentLevelNo)2	1.074	1.041	1.108	<0.001
factor(currentLevelNo)3	1.028	0.982	1.077	0.234
factor(currentLevelNo)4	1.053	1.011	1.097	0.013
factor(currentLevelNo)5	0.916	0.805	1.043	0.184
avgAccuracy	1.005	1.004	1.006	<0.001
currentAccuracy	0.995	0.995	0.996	<0.001
bestTime	1.241	1.232	1.25	<0.001

Since we used log transformation on the response variable, we transformed our model's coefficient of estimate exponentially. We cound interpret the intercept of our model output as the average log reaction time for players who gave the wrong answer on question 0.1, without guide enabled, who was on level 1, whose average and current accuracy was zero, and best time was zero, was expected to respond in 1.612 seconds. On average, the questions with the correct response took 1.13 times longer than the questions that were answered wrong. We could interpret each of the current question factor as its average response time ratio between question one and the selected question, controlling for other variables. For instance, we could interpret question 0.13's parameter as on average, the adjusted response time for answering this question was 1.035 times of the average time to answer question 0.1. The expected response time for questions with guide enabled was 4.9%

higher than the questions without, controlling for other variables. For the each of the level parameter, we could interpret them as their adjusted time ratio in relation to the average response time for level 1 questions. For example, the expected response time for level 5 questions was 0.916 times of the expected response time for level 1 questions, controlling for other variables. Every percentage increase in player's average accuracy, we were expecting a 0.5% increase in his or her adjusted response time, and for every percentage increase in the player's current accuracy, we were expecting a 0.5% decrease in his or her adjusted response time. For every second increase in player's best time, the expected response time for this player increase by 24.1%, controlling for other variables.

Note that it would be reasonable to treat easiness as a numeric variable, but nonsensical to treat resp and CurrentQuestion as numeric variables. If we were to treat current question as a numeric variable, we would be assuming that there was an ordinal relationship between each question in terms of reaction time. However, our previous plot disproved that, as the reaction time for all questions appeared to be uncorrelated. It would be unnecessary to treat resp as a numeric variable since it was a binary variable with two possible outcomes. We could treat easiness as a numeric variable because it was a continuous measure with ordinal relationship between the variables – the higher the value the easier the question.

Ordinary linear regression for player effects

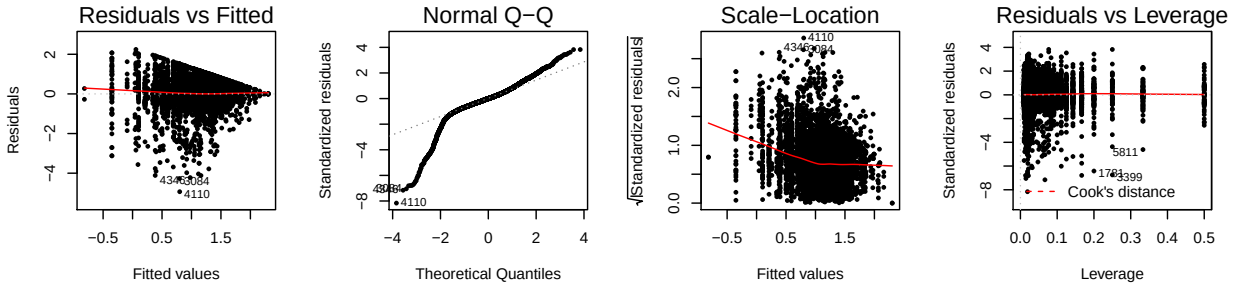


Figure 4: Diagnostics for Linear Model of lrt ~ SID

The diagnostics plot of our model suggested several minor violations in linear model residual assumptions. Scale location plot showed that the residual variance did not appear to be constant. Besides the line of cutoff, it appeared that the zero expectation assumption was mostly met, and our residual appeared to be independent. The normality assumption was not fully met, and it had tails on both ends. We also observed couple potential outliers (point 4410, 4346, 3084, 1781, 3399 and 5811). We could interpret our model's parameter as the expect log response time for each player, which in this case was estimated by the average log response time. For example, the expected log response time for player 161461 was 0.85121.

Figure 5 provided us with insights with regard to how the expected log response time of the subjects was related to other variables. We could see that there does not seem to be a relationship between player's fluency and his or her log response time. From the plot between the random effects from the muti-level model (where we used current question as fixed effect and player ID as varying intercept, and correctness as response variable) and the expected player log response time, we could see a possible pattern. This suggested that there may exist a relationship between the average log-reaction time of each player and this player's ability to answer the question correctly, controlling for the type of question. The expected log reaction time and the expected log average reaction time for each player showed some pattern with points drifting below the y=log(x) and the y=x line. This was caused by the smaller number of questions being answered by each player, thus making the two averages appeared closer on the plot. We still can conclude that the average log reaction time for each player was not the same as the log average reaction time.

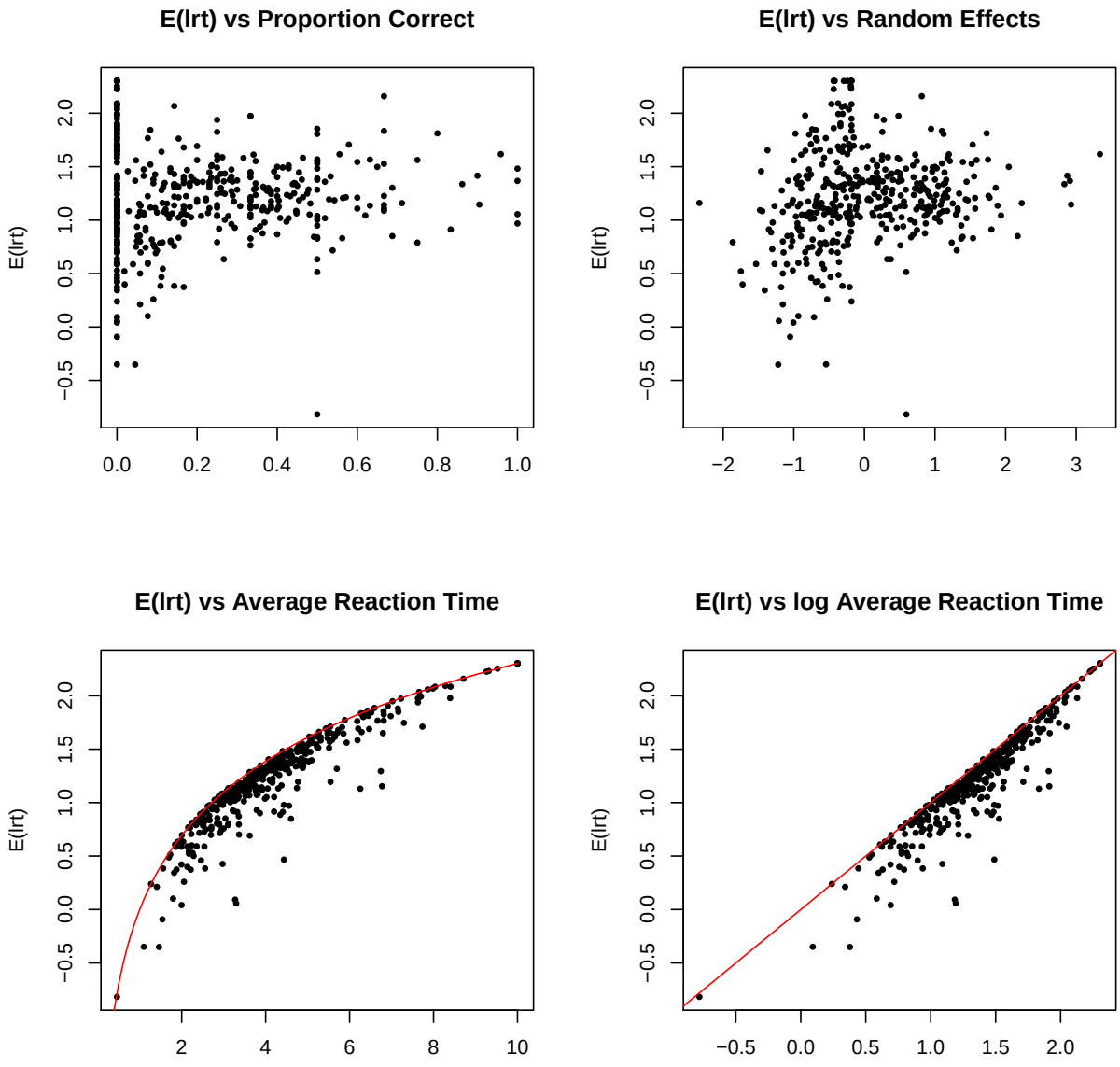
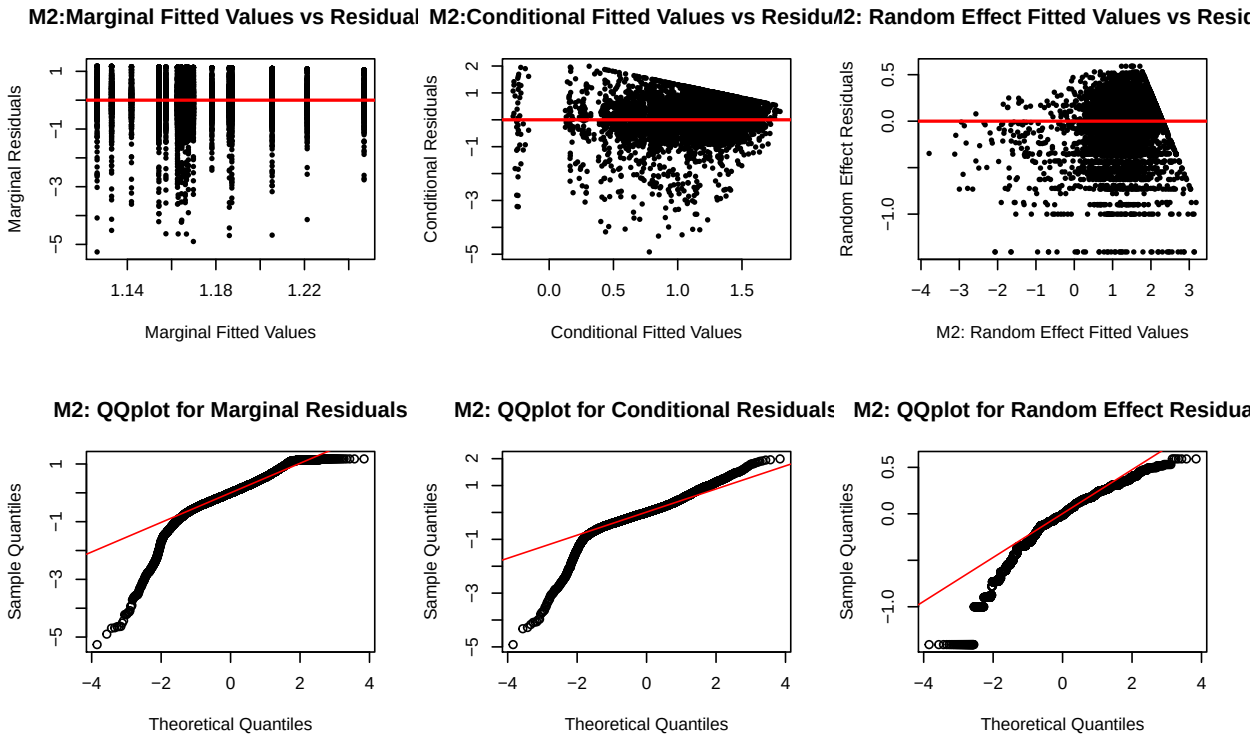


Figure 5: Expected Log-Reaction Time vs player fluency, random effects from the glm model, average reaction time and log of the average reaction time, of the players

Mixed Effects Models



Our mix model's marginal and conational residuals appeared to have zero expectation and constant variance, suggesting that the model was an appropriate fit on the data. Our normality assumption was not fully met, with tails on both end of the normal line. The fixed effect of our model represented the expected log-response time from all subjects at each questions, taking the difference between individuals into account. For example, we conclude that the expected log reaction time for players to answer question 0.1 is 1.186 when grouping by individuals. The random effect of our model represented the effect of individuals on the log reaction time on the question. For example, individual 164461's random effect suggested that when controlling for the question, the subject was expected to have 0.3022 seconds shorter log reaction time than the average log reaction time across all players.

Our model suggested that the effect of questions on players' log response time was higher than the average log response time on each question across all players. This meant that the effect of the question was larger than what it appeared from the unbiased estimator in our linear model, and that the individual effects played a less important role in estimating the log response time. Our model's random effect suggested that in almost all cases, the random effect had less influence on the log reaction time than the average log reaction time on each player across all players. We also observed a pattern in the random effects, suggesting that we may require more information in our model to capture the underlying reason. For subjects with smaller average log reaction time, it appeared that their random effects were closer to the average, and for subjects with larger log reaction time, their random effect appeared to be more different than the average.

We used back-fit fixed effects and forward-fit random effects of an LMER model (from the LMERConvenience-Functions package in R) to search for the best model under the BIC criteria. We started with the largest model: log reaction time as the response variable, with whether the player responded the question correctly, the current question, whether guide is enabled, the current level, average accuracy, total star count, accuracy of the current session, items player, and the best time of the player as fixed effects, and we used individual players as random effects. We then compare the full model, along with the searched result, with our smallest model, which was a mixed effect model with no fixed effects and individual players as random effect.

From our anova table, we can see that under BIC criteria, the model selected by the automatic model selection

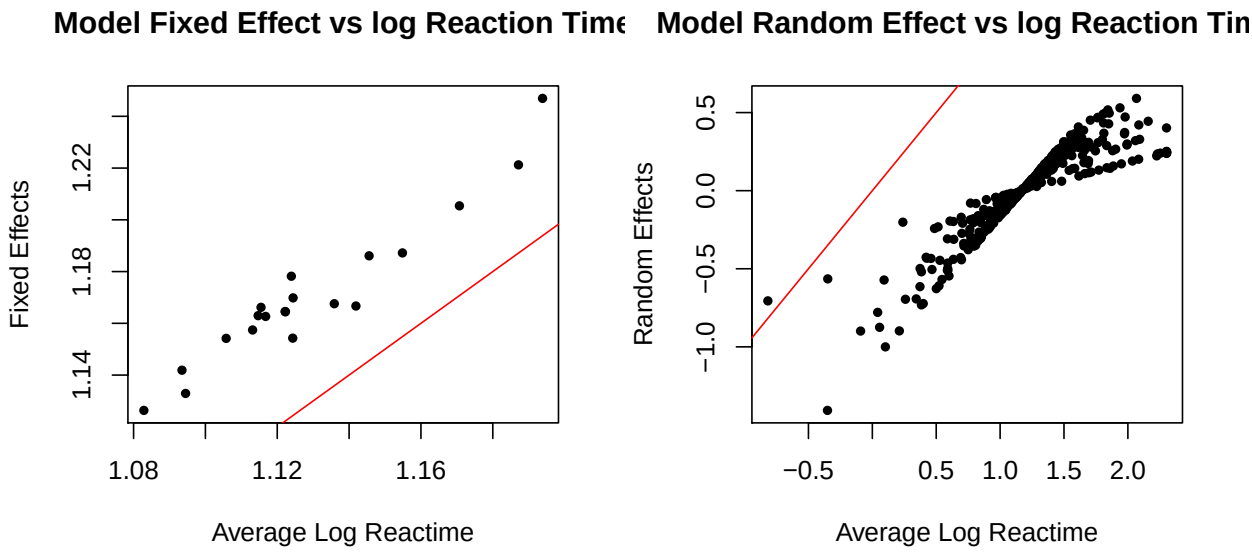


Figure 6: Mixed Model Parameters vs Data Average

has the best BIC, and the AIC of the model came close with the initial full model, but still with considerable improvement. Therefore, we selected the model with whether the player responded the question correctly, current level, average accuracy, current accuracy and player best time as fixed effects, and individual players as random effects.

Table 3: Data: game.data

	Df	AIC	BIC	logLik	deviance	Chisq	Chi Df	Pr(>Chisq)
lm.q4.small	3	15933	15954	-7964	15927	NA	NA	NA
bic_best.1	11	13550	13627	-6764	13528	2399	8	0
lm.q4.mix	22	15956	16111	-7956	15912	0	11	1
lm.q4.full	33	13553	13785	-6744	13487	2425	11	0

Our model's fixed effects suggested that when considering individual players' difference, the response time was expected to be 10% higher when the question was answered correctly. Each of the current level factor corresponded to the difference in response time with that of the current level one's. For example, we expect that it would take on average 9.6% longer to respond level 2 questions than to respond level 1 questions. For each additional percentage in player accuracy (in percentage), we expected an average 0.38% increase in response time, and for each additional percentage in current accuracy, we expected the average response time to be 0.46% shorter. For each additional second in player's best time, we expect that player's response time to be 20.1% higher, when taking individual players' average response time into account. The model's random effects summarized the effects of individual player's response time on the response time of the question. For instance, subject 161461 on average had a 11.71% less response time on questions, controlling for the variables that we had discussed in the fixed effects.

Models combining reaction time and correctness of response

In our modeling in the previous part, we concluded that there existed a relationship between the scaled reaction time and the probability of answering correctly. In particular, for every second increase in player reaction time, the average probability of a player getting the correct answer increases by 0.12 percent. This was reinforced in our current analysis when we modeled the log reaction time as the response variable. The mixed model results suggested that having a correct response was associated with a longer reaction time, and for players with higher average accuracy, their response time was also expected to be longer. Thus, we concluded that it is entirely possible that the correctness of response was related to the reaction time in some way.

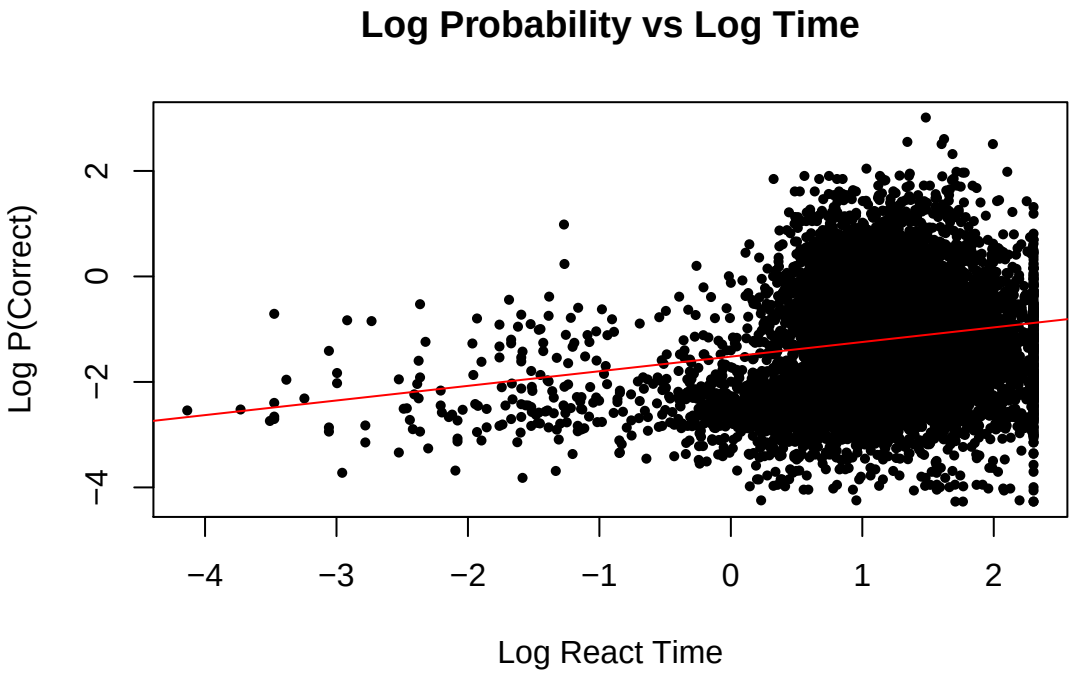


Figure 7: From the figure we can see that log probability may be a reasonable predictor of log response time

The plot suggested that log probability of correctness may be a good predictor of the log response time on the questions, as the two were hypotheized to be positivey corrlated. We attempted to expand the model by including log probability, correct response, current question, whether guide was enabled, the current level number, the average accuracy of the subject at the time of answer, the total star count, the current accuracy of the subject, number of items played, and the best time as covariates. After stepwise AIC selection, we decided to retain our full model. However, we had some concerns with our model fitness from our diagnostics plot. We can observe that the variance was not constant and the normal residual assumption was not quite met.

Level 1: $lrt_i = \beta_1 * (Question_i + \alpha_{[j]i}) * \rho + \epsilon^2$

where $\sigma^2 \sim N(0,1)$, $\rho \sim \sigma^2 \sim N(0,\sigma^2)$, $\rho \sim N(0,1)$ and $\sum Question = 0$

Level 2: $\alpha_j = \beta_0 + \tau^2$ where $\tau^2 \sim N(0,\eta^2)$

Upon inspecting both simulation results, we can see that the model fitted under joint.mcmc.1 appeared to be better than joint.mcmc.0. All the parameters converged to stationary distribution and the autocorrelation appeared to satisfy the markov assumption of independence. This was not achieved by joint.mcmc.0. Not

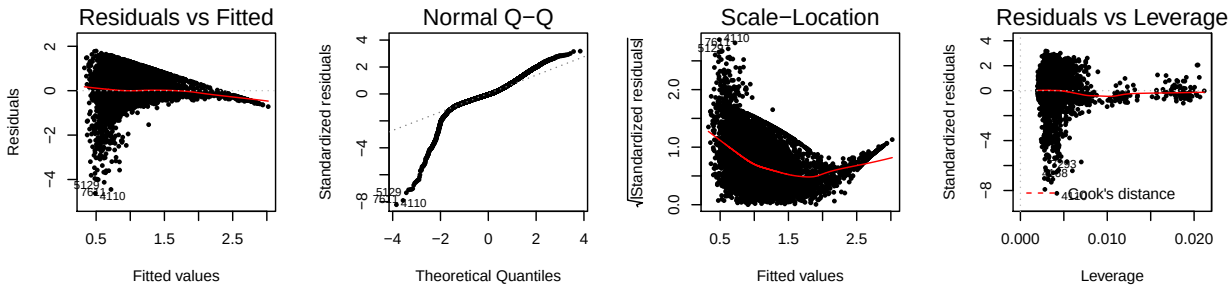


Figure 8: Diagnostics for Auto Selected Linear model

only were the rhats high on the parameters, suggesting non-convergence, the autocorrelation plot showed poor fit under the markov chain assumption. Therefore, the results in joint.mcmc.0 was not reliable.

It appeared that joint.mcmc.1 was set up with more prior assumption in the residual distributions than joint.mcmc.0. In particular, model 1 drew the residuals based on the 95% confidence interval of the same residuals after simulation in model 0. Therefore, model 1 may have been better because it was improved from model zero by drawing the simulated results, this creating a better initial state that lead to convergence. Similar to burning off the previous data.

The parameters of the markov chain suggested that the probability of answering the question correct did have a relationship with the response time that it took. In particular, the relationship was positively correlated. Thus, our simulated result matches with the conclusion that we reached in part b.

Appendix I

Gelman, A. & Hill, J. (2007). Data Analysis Using Regression and Multilevel/Hierarchical Models. NY: Cambridge Univ Press.

Lomas, D., Patel, K., Forlizzi, J.L., and Koedinger, K.R. (2013). Optimizing Challenge in an Educational Game Using Large-Scale Design Experiments. Paper presented at CHI 2013, Paris, France. Obtained online at http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.480.2493&rep=rep1&type=pdf

Lynch, Scott M. (2007). Introduction to Applied Bayesian Statistics and Estimation for Social Scientists. New York: Springer.