The Statistics Behind Perceptual Decision Making

Wei Dou, Jacobo Pereira-Pacheco and Yu Zhu

Abstract

The current study aims to explore the relationship between response time, evidence strength level, accuracy, and confidence in a Two-Choice Decision (TCD) framework. We fit a randomized block model to understand the relationship between evidence strength levels and response time. The results show that stronger evidence is associated with faster response time. We also fit two mixed effect logistic regression models with accuracy and confidence as the response variables respectively. Both models involve the same predictors. Through inference, we find that at certain levels of evidence strength, faster decisions are related to a decrease in the odds of an accurate TCD and high level of confidence. Whereas, keeping response time constant, stronger evidence is more likely to result in an accurate decision and higher confidence. To test and explain the performance of our models, we make predictions for accuracy and confidence with measured good performance. Our results support the drift diffusion model which assumes that the decision formation is based on the accumulated evidence provided by sensory stimuli.

KEY WORDS: Randomized Block Model, Mixed Effects Logistic Regression Model, Prediction.

1 Introduction

Perceptual decision making is a term to describe how sensory information is used to guide behavior toward the external world. The drift diffusion model (DDM) [1] has been developed to provide an account of accuracy, response time, and confidence occurring in two-choice perceptual decisions. DDM assumes that decisions are made by extracting sensory evidence from the stimulus and subsequently accumulating the evidence over time. Once the amount of accumulated evidence reaches one of two response thresholds, a response is elicited. When the sensory evidence is stronger, it is accumulated faster and more rapidly hits the decision boundary. It has been also proposed that stronger evidence leads to higher confidence, compared to weak evidence [2].

Dou and Samaha [3] examined how evidence strength influences confidence in observers' perceptual decisions using a two-choice discrimination task with six motion coherence (evidence strength) levels. In our current project, we will explore the behavioral data from their study to look at if observers' actual responses follow the predictions of DDM.

1.1 The Experiment

In the experiment, participants were presented with several moving dots, with some moving horizontally and some moving randomly. Participants pressed buttons to indicate the decision of the motion direction of horizontally moving dots, and their confidence in their decision. There were six types of stimuli with different levels of coherence. Stimulus with higher coherence levels provided stronger evidence. So the experiment is a repeated measurement design with one factor (evidence strength). Each participant completed 1080 trials in total, with 180 trials at each strength level of stimulus. The presentation of the six stimulus were fully randomized.

1.2 The Data

The data is from a repeated measured experiment with evidence strength as the factor variable. The response variables include response time, accuracy, and confidence of the decision. Each of the 25 participants completed 180 trials on each strength level, amassing to a total of 1080 theoretical trials. However, the behavioral data has been trimmed by the experimenters due to excluding trials in which the EEG signals in the experiment were noisy. The details of the five variables are listed below:

- **Participant Number (P):** Integer variable, the index for the participants
- **Strength Level (S):** Ordinal variable with 6 levels: $1\%, 4.5\%, 8\%, 12\%, 25\%$ and 40%
- **Response Time (R):** Numeric variable
- **Accuracy (A):** Binary variable with 0 and 1, 0 represents inaccurate TCD and 1 represents accurate

TCD
- **Confidence (C):** Ordinal variable with 4 levels: 1, 2, 3 and 4

1.3 The Goal

The central question that we want to address is how sensory evidence strength statistically influences decision making according to the DDM. To answer this central question, we will regard each trial from participants as an observation, and explore three technical aspects of the data: (1) what are the statistical relationships between response time, evidence strength level, accuracy, and confidence that we can further uncover; (2) whether confidence can be predicted by evidence strength and response time; (3) whether accuracy can be predicted by evidence strength and response time. From these results we will then be able to qualitatively and quantitatively examine how evidence strength influences decision making in this psychological experiment.

1.4 Prior Analysis

Dou and Samaha [3] performed three one-way ANOVAs to participants' means of the behavioral data to examine how evidence strength affects accuracy, response time, and confidence respectively. They found that both accuracy and confidence increases with stronger evidence, and response time decreases with stronger evidence.

2 Exploratory Data Analysis

2.1 Missing Value and Influential Data Analysis

There is no missing value in our data set. And the histogram of the continuous variable response time is right-skewed with potential indication of no outliers or influential observations. Based on these findings, there is no need to do data imputation or transformation before fitting the models. And we can do further analysis for possible outliers and perform diagnostics with the residual plots after the model establishment.

2.2 Variable Explorations

To discern the basic relationships between the variables, box plots and mosaic plots are helpful. We first utilize side-by-side box plots for response time versus evidence strength, participants, confidence, and accuracy respectively.

- **Response Time vs Evidence Strength:** The mean and variability in response time decrease with the increase of strength level.
- **Response Time vs Participants:** No obvious pattern in the box plots. The mean and variability in response time may not depend on the participants.
- **Response Time vs Accuracy:** The mean and variability of the response time decrease when the TCD is accurate.
- **Response Time vs Confidence:** The mean and variability of the response time decrease with the increase of the confidence level.

Then we apply the mosaic plots for the evidence strength versus confidence and accuracy respectively.

- **Evidence Strength vs Accuracy:** As the evidence strength level increases, we observe higher proportion of the accurate TCD, which suggests the positive relationship between evidence strength and accuracy.
- **Evidence Strength vs Confidence:** As the evidence strength level increases, we observe higher proportion of the higher confidence level, which suggests the positive relationship between evidence strength and confidence.

We also perform a Chi-square tests for the independence between evidence strength and confidence, and evidence strength and accuracy. Both of the p-values are approximately equal to 0, so we can reject the null hypotheses that the evidence strength is independent of confidence, and that the evidence strength is independent of accuracy, respectively. We can conclude that there exists significant positive relationships between evidence strength and confidence, as well as between evidence strength and accuracy.

2.3 Interactions between Response Time and Evidence Strength

In order to check if there are significant interaction effects between the response time and evidence strength, interaction plots are used to examine the interactions based on the tentative fitting of two logistic regression models as shown in Figure 1.

The first logistic regression model is predicting accuracy with response time, evidence strength and the interaction term between response time and evidence strength. For the plot of predicted log odds of TCD being accurate versus response time for each strength level, the slopes of lines for strength level 25% and 40% are different from

the other four levels. They also intersect with each other. This interaction plot suggests the interaction effect exists between evidence strength level and response time for predicting accuracy.

The second logistic regression model is predicting confidence with the same predictors. Here we map the confidence into 2 different categories: confidence levels of 0 and 1 are combined into the category of 0, and confidence levels of 3 and 4 are combined into the category of 1. Different from the interaction plot for accuracy, the lines in the plot of predicted log odds of confidence under category 1 versus response time for each strength level are mostly parallel with each other, except for strength level 40%. So we can still see the existence of the interaction effects between evidence strength level and response time for predicting confidence when confidence is treated as a binary variable.

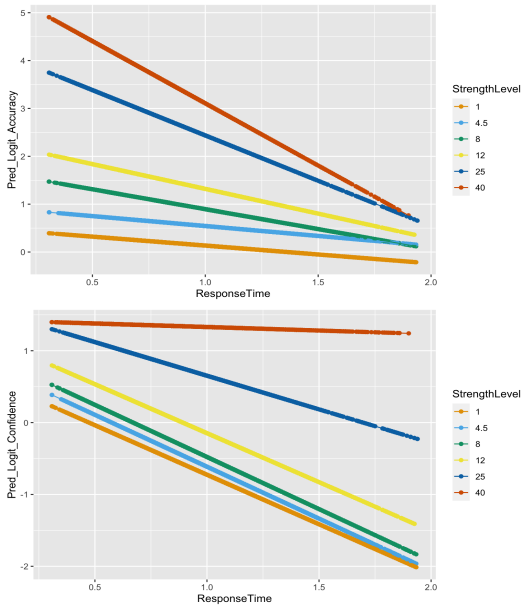

Figure 1: The interaction between response time and evidence strength

3 Randomized Block Design

Our first model of interest is a randomized block model (1) where y_{ij} is the response time, τ_i is the evidence strength for $i = 1, ..., 6$, μ is the grand mean, β_j is the blocking factor for participant $j = 1, ..., 25$, and ϵ_{ij} is the random error normally distributed with constant variance.

$$y_{ij} = \mu + \tau_i + \beta_j + \epsilon_{ij} \quad (1)$$

$$\epsilon_{ij} \sim N(0, \sigma_s^2)$$

Our interest in this model is to further understand the statistical relationship between response time and evidence strength while controlling for the variation of different participants in the experiment.

3.1 Diagnostics

In order to verify our model assumptions, we have to take into account diagnostics about the assumptions of our randomized block model which are homoskedasticity and normality with regards to the residuals. After investigation of the residual plots of our model, we decided that the model may benefit from a log transformation due to a moderate departure of the normality assumptions with respect to the residuals. After transforming our response variable via a logarithm transformation, the Q-Q plot was more stable across the Q-Q line (Figure 2), and we decided to move forward with the log-transformed model. Homoskedasticity of the residuals was present in both forms of the model.

3.2 Model Hypothesis Testing

The null hypotheses here are (1) $H_0 : \tau_i = 0$ for all i, and (2) $H_0 : \beta_j = 0$ for all j. After fitting our randomized block model we retrieve a p-value that is ≈ 0 for both evidence strength and participant. Considering a significance level of $\alpha = 0.05$, we are able to reject both null hypotheses in favor of the alternative. Hence, considering the treatment, we can state that there are significant differences in means of log-transformed response time between the evidence strength levels. And considering the block factor, we can state that there are significance differences in means of log-transformed response time between participants.

3.3 Analysis and Interpretation

Of particular interest is to further understand which evidence strength levels differ in log-transformed response time. Using Tukey's Honest Significant Difference (Tukey's HSD) to examine pairwise comparisons under a 95% confidence interval we conclude that out of 15 total pairwise comparisons, 14 exhibited statistically significant differences between means. Specifically, except for, '4.5%-1%', all the confidence intervals for the pairwise difference are significantly different from 0 (the confidence intervals do not contain 0). As in Figure 3, the upper bounds of all significant confidence intervals are negative, and can be interpreted such that the mean of log-transformed response time for the evidence

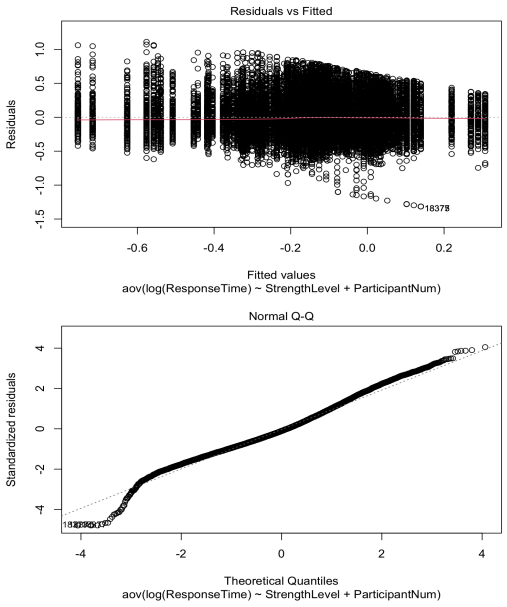

Figure 2: Residuals plots for the randomized block model

strength level of 40% is significantly smaller than 25% ($\mu_{RT:40} < \mu_{RT:25}$). Using that same notation and analyses, we conclude that the same analyses is applicable to the following pairs: $\mu_{RT:25} < \mu_{RT:12}$, $\mu_{RT:12} < \mu_{RT:8}$, and $\mu_{RT:8} < \mu_{RT:1}$.

As our data set indicates slight differences for the sample size under each factor level, we are also interested in applying Schéffe's method. The result is similar to but more strict than the Tukey's HSD method, we have 11 out of 15 contrasts that are significant under a 95% confidence interval. Schéffe's method indicates that in addition to '4.5%-1%' from Tukey HSD, we also also have '8%-1%', '8%-4.5%', '40%-25%' as pairs that do not exhibit a statistically significant difference in means.

Overall, we can conclude that as the evidence strength increases, the response time becomes shorter, which confirms the expectation of the relationship between evidence strength and response time in EDA.

4 Train-test Split

With respect to our mixed effect logistic regression models, we are interested in using the data from the experiment to infer and predict statistical information about accuracy and confidence related to general observations. Thus, we will be firstly partitioning our data into two subsets: the training set and the testing set. Therefore, our training set is formed by random sampling of 80 observations under each of the combinations of strength level and participants. In this case, the size of the training set

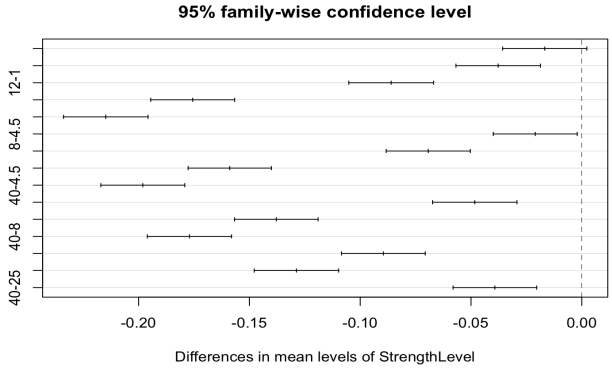

Figure 3: Pairwise comparisons for evidence strength under Tukey's HSD

is 12,000, which is about 60% of the population. The remainder of about 40% is assigned to the testing set. From the box plots of the response time of two subsets as well as the empirical cumulative distribution functions (Figure 4), the distributions of the response time in the training and testing sets are of high similarity.

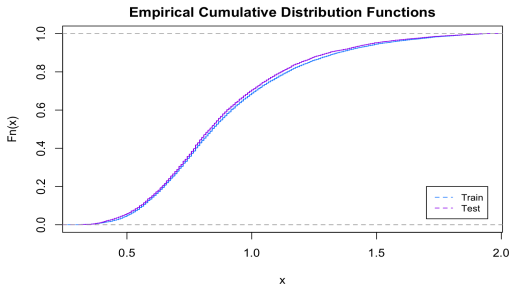

Figure 4: Similar distributions between the training and testing sets

Since our research question with respect to accuracy and confidence is inference and prediction of any random individual using the experimental data, we will first build our logistic regression model based on the training data, use the training model to solidify a general understanding of the variable relationships, and then use the test data set to predict accuracy and confidence respectively, and test model performance.

5 The Accuracy Model

Our second modeling framework is the construction of two mixed effect logistic regression models to infer and predict accuracy: 1) the full model M1 (2) with a fixed effect evidence strength (X_S), continuous response time variable (X_R), the interaction term between evidence strength and response time ($X_S X_R$), and random effect

participants (S_P); 2) the additive model M2 (3) with a fixed effect strength level (X_S), continuous response time variable (X_R), and random effect participants (S_P). P_A is defined as the vector of probability that the TCD is accurate.

$$logit(P_A) = \beta_0 + \beta_S X_S + \beta_R X_R + \beta_{SR} X_S X_R + S_P \quad (2)$$

$$logit(P_A) = \beta_0 + \beta_S X_S + \beta_R X_R + S_P \quad (3)$$
$$S_P \sim N(0, \sigma_s^2)$$

5.1 Model Selection

We begin with our full model M1 which includes the interaction and also fit our additive model M2 without the interaction. We then perform a Likelihood Ratio test (LR test) to examine the goodness-of-fit of our models as well as to indicate which model fits the data better. This is plausible since M2 is a nested model within M1. For the LR test, our null hypothesis is that $H_0 : \beta_{SR} = 0$, or more generally, the additive model M2 fits the data better. The resulting p-value is ≈ 0 which is < 0.05, indicating that we can reject the null hypothesis and conclude that the full model M1 is preferred. Thus, we choose M1 as our final mixed effect logistic regression model for accuracy.

5.2 Diagnostics

Through the Pearson's residuals plot and Deviance residuals plot in Figure 5, they present the patterns of skewness with large proportion of fitted values (predicted log odds of TCD being accurate) greater than 0. One of the reasons is because the response variable accuracy is imbalanced with only 25% of observations being inaccurate for the TCD. In order to make our model more valid, for future work we can use some re-sampling methods such as up or down sampling to make the variable accuracy more balanced, or apply the modified models such as weighted logistic regression model. In addition we can examine the residuals of our accuracy model via a binned residual plot, in which the average residual is plotted versus the average fitted value for each bin. Most observations fall within the standard error bounds which indicates an adequate fit, however when fitted values are smaller than 0.7, outliers begin to appear.

5.3 Model Hypothesis Testing

To further understand the relationship between the predictor variables and the probability of accuracy, we want

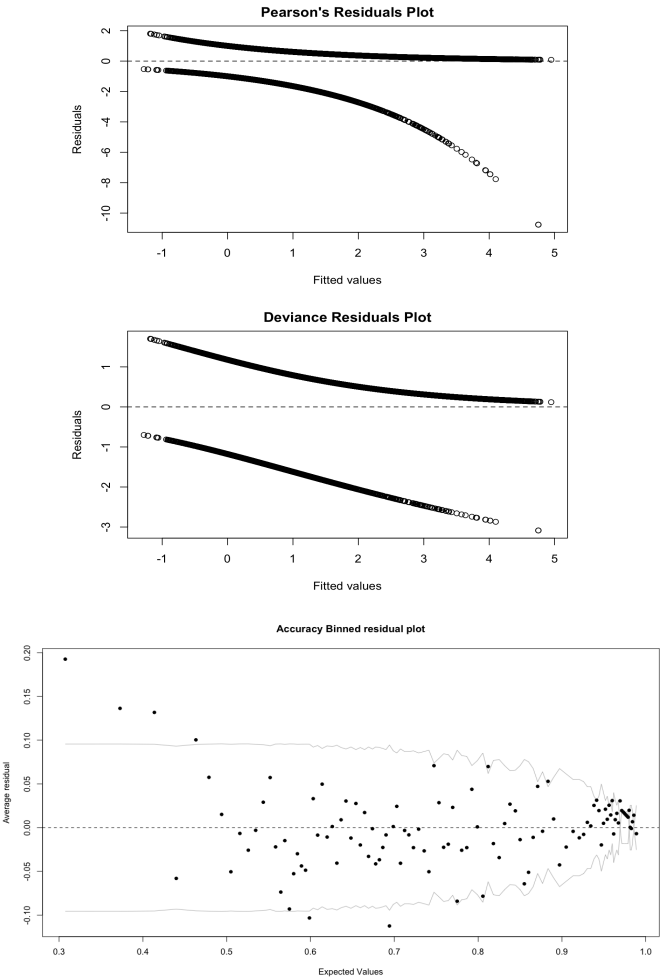

Figure 5: Residuals plots for the Accuracy model

to look at the p-value of individual variables from the results given by Table 1 to confirm any present associations in our experiment. Evidence strength is a categorical variable and so the baseline is related to level 1%, and level 4.5%, 8%, 12%, 25%, and 40% are related to the baseline. Included in our model is also our continuous variable response time, as well as subsequent interaction terms between response time and the aforementioned evidence strength.

Interpreting p-values for each of these variables, we have strength level 4.5%, 8%, 12%, 25%, and 40% all with respective p-values of 0.045, $9.89e^{-02}$, $< 2e^{-16}$, $< 2e^{-16}$, $< 2e^{-16}$. The preceding p-values are less than 0.05, indicating that we can reject the null hypothesis of $\beta_{4.5} = 0, \beta_8 = 0, \beta_{12} = 0. \beta_{25} = 0, \beta_{40} = 0$ and conclude that factor evidence strength is significant under 0.05 a significance level.

Next we have response time, with a p-value of $0.0006 < 0.05$, so we can reject the null hypothesis of

$\beta_R = 0$ and conclude that there exists a significant linear relationship between the log odds of a TCD being accurate and response time.

Finally, our interaction terms are pairs consisting of strength level and response time described as such, 4.5:R, 8:R, 12:R, 25:R, 40:R; with p-values of 0.974, 0.045, 0.0003, $5.85e^{-05}$, $1.45e^{-10}$ respectively. The p-values for 8:R, 12:R, 25:R, and 40:R are all < 0.05, thus we can reject their null hypothesis of $\beta_{8R} = 0$, $\beta_{12R} = 0$, $\beta_{25R} = 0$, $\beta_{40R} = 0$ respectively and conclude that there are interactions between evidence strength and response time that are significant under a significance level 0.05.

5.4 Model Interpretation

From the model output, we firstly notice how the coefficients for strength levels 4.5%, 8%, 12%, 25% and 40% are all positive and increase in value as the strength level becomes larger. This is an interesting pairing when one looks at the interaction terms between strength level and response time, in which the coefficient becomes smaller in interaction effect as strength level becomes larger. To quantify this result we can calculate the change in odds of a TCD being accurate for each strength level given a 1-unit increase in response time via each respective interaction term:

- At a strength level of 1%, a 1-unit increase in response time is associated with an estimated 40.15% decrease in the odds of TCD being accurate.
- At a strength level of 4.5%, a 1-unit increase in response time is associated with an estimated 40.55% decrease in the odds of TCD being accurate.
- At a strength level as 8%, a 1-unit increase in response time is associated with an estimated 60.88% decrease in the odds of TCD being accurate.
- At a strength level as 12%, a 1-unit increase in response time is associated with an estimated 73.42% decrease in the odds of TCD being accurate.
- At a strength level as 25%, a 1-unit increase in response time is associated with an estimated 81.68% decrease in the odds of TCD being accurate.
- At a strength level as 40%, a 1-unit increase in response time is associated with an estimated 86.76% decrease in the odds of TCD being accurate.

We also compare the odds of TCD being accurate under different strength levels with a constant response time 1.0:

- Holding the response time constant at 1.0, the odds of TCD being accurate for strength level 4.5% is estimated to be 1.53 times that of TCD being accurate for strength level 1%.
- Holding the response time constant at 1.0, the odds of TCD being accurate for strength level 8% is estimated to be 2.10 times that of TCD being accurate for strength level 1.
- Holding the response time constant at 1.0, the odds of TCD being accurate for strength level 12% is estimated to be 3.46 times that of TCD being accurate for strength level 1%.
- Holding the response time constant at 1.0, the odds of TCD being accurate for strength level 25% is estimated to be 9.64 times that of TCD being accurate for strength level 1%.
- Holding the response time constant at 1.0, the odds of TCD being accurate for strength level 40% is estimated to be 19.94 times that of TCD being accurate for strength level 1%.

5.5 Model Prediction

We set our threshold as 0.5. The test error of our model is about 0.23, so the model M1 overall correctly predicted the accuracy of a TCD 77% of the time. Based on the confusion matrix (Figure 6), our model correctly predicted that the TCD will be accurate on 6264 observations and inaccurate on 234 observations.

Additionally, the area under ROC curve (AUROC) of our model (Figure 7) is 0.768, which is overall acceptable but can be improved. We also have a pretty small specificity as 0.122 and a large sensitivity as 0.959. Specifically, we can correctly predict about 12.2% inaccurate TCD of all the inaccurate TCD; and we can correctly predict about 95.9% accurate TCD of all the accurate TCD. This result is under our expectation due to the imbalanced binary classification as mentioned in the diagnostics section. For better prediction performance, we need to modify our threshold by applying cross-validation or to make our data balanced.

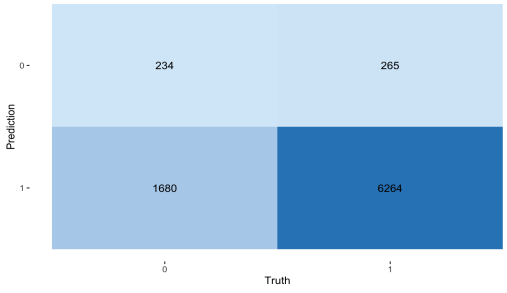

Figure 6: The confusion matrix for the Accuracy model

Table 1: Mixed Effects Logistic Regression Models for Accuracy

	Dependent variable:	
	Accuracy	
	Full Model M1	Additive Model M2
StrengthLevel4.5	0.432** (0.010, 0.854)	0.428*** (0.299, 0.557)
StrengthLevel8	1.168*** (0.738, 1.598)	0.752*** (0.619, 0.885)
StrengthLevel12	2.053*** (1.599, 2.507)	1.266*** (1.122, 1.411)
StrengthLevel25	3.450*** (2.862, 4.037)	2.335*** (2.142, 2.528)
StrengthLevel40	4.501*** (3.715, 5.287)	3.112*** (2.853, 3.371)
ResponseTime	−0.513*** (−0.807, −0.220)	−0.903*** (−1.064, −0.742)
StrengthLevel4.5:ResponseTime	−0.007 (−0.416, 0.402)	
StrengthLevel8:ResponseTime	−0.425** (−0.842, −0.009)	
StrengthLevel12:ResponseTime	−0.812*** (−1.253, −0.371)	
StrengthLevel25:ResponseTime	−1.184*** (−1.761, −0.606)	
StrengthLevel40:ResponseTime	−1.509*** (−2.285, −0.732)	
Constant	0.650*** (0.317, 0.983)	1.033*** (0.801, 1.265)
Observations	12,000	12,000
Log Likelihood	−5,541.612	−5,560.464
Akaike Inf. Crit.	11,109.220	11,136.930
Bayesian Inf. Crit.	11,205.330	11,196.070
Note:		*p<0.1; **p<0.05; ***p<0.01

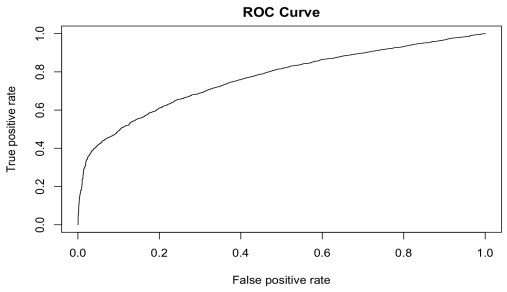

Figure 7: The ROC curve for the Accuracy model

6 The Confidence Model

As we are interested in predicting the confidence under a more general division: high confidence level and low confidence level, we first propose to map the four levels of confidence into two: as 1 if the level of confidence is 3 or 4, and as 0 if the level is 0 or 1. Essentially, we treat confidence as a binary variable, with category of 1 suggesting a high level of confidence and 0 for a low level. Similarly, we perform two logistic regression models for prediction: 1) the full model M3 (4) with a fixed effect evidence strength (X_S), continuous response time variable (X_R), the interaction term between evidence strength and response time ($X_S X_R$), and random effect participants (S_P); 2) the additive model M4 (5) with a fixed effect strength level (X_S), continuous response time variable (X_R), and random effect participants (S_P). P_C is defined as the vector of probability that the confidence level is high (level of 3 or 4).

$$logit(P_C) = \beta_0 + \beta_S X_S + \beta_R X_R + \beta_{SR} X_S X_R + S_P \quad (4)$$

$$logit(P_C) = \beta_0 + \beta_S X_S + \beta_R X_R + S_P \quad (5)$$

$$S_P \sim N(0, \sigma_s^2)$$

Here, we apply the same training and testing set as the previous logistic model.

6.1 Model Selection

We begin with our full model M3 which includes the interaction term and also fit our additive model M4 without the interaction term. We then perform a LR test to examine the goodness-of-fit of our models and make comparison between M3 and M4 since M4 is nested in M3. The null hypothesis is $\beta_{SR} = 0$, or more generally, M4 fits model better. Since the p-value is $0.00048 < 0.05$, we can reject the null hypothesis and the LR test indicates the full model M3 is a better fit. Hence, we choose the full model M3 as our final logistic regression model for confidence.

6.2 Diagnostics

Through the Pearson's residuals plot and Deviance residuals plot in Figure 8, we can observe that our model fits appropriately as there is symmetry among the residuals and no indication of influential observations or skewness. Additionally from the binned residuals plot, there is no trend of outliers and almost all of the observations are within the standard error bounds indicating that the model is a good fit for the data. Overall, we can conclude that the assumptions of our model are valid under the diagnostics of the residual plots.

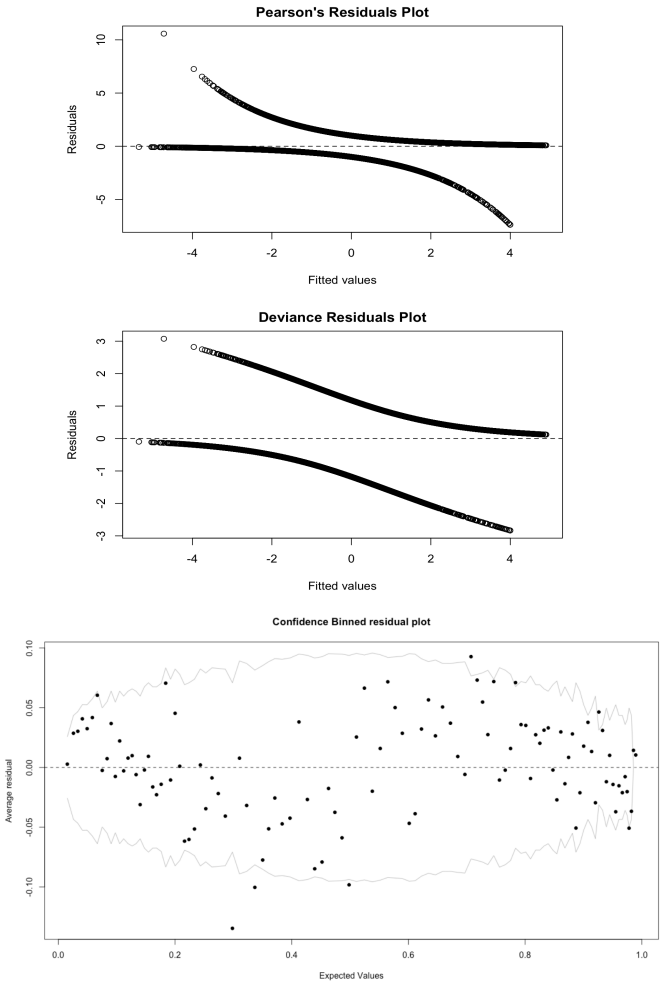

Figure 8: Residuals plots for the Confidence model

6.3 Model Hypothesis Testing

Referring to Table 2, we first look at strength level 4.5%, 8%, 12%, 25% and 40% with respective p-values of 0.133, 0.012, 0.001, 4.31e-08 and 6.05e-09. Except for the p-value of strength level 4.5 as $0.133 > 0.05$, all of the p-values left are less than 0.05. So we can reject each of the null hypothesis of $\beta_8 = 0$, $\beta_{12} = 0$, $\beta_{25} = 0$ and $\beta_{40} = 0$ and conclude that the factor evidence strength is significant under 0.05 significance level.

Next we have response time, with a p-value $<$ 2e-16. Hence, we can reject the null hypothesis of $\beta_R = 0$ and conclude that the variable response time is significant under 0.05 significance level.

As for the interaction between evidence strength and response time, The p-value for 40:R is < 0.05, thus we can reject the null hypothesis of $\beta_{40R} = 0$ and conclude that there is an interaction between evidence strength and response time that is significant under significance level 0.05.

6.4 Model Interpretation

Similar to the M1 for accuracy, we are interested in the change for the odds of confidence in high level when we adjust the response time and evidence strength. To illustrate, we first calculate the change in odds of confidence in high level under each constant strength levels when the response time increases for 1 unit via the respective interaction terms:

- At a strength level of 1%, a 1-unit increase in response time is associated with an estimated 90.40% decrease in the odds of confidence in high level.
- At a strength level of 4.5%, a 1-unit increase in response time is associated with an estimated 91.95% decrease in the odds of confidence in high level.
- At a strength level of 8%, a 1-unit increase in response time is associated with an estimated 93.06% decrease in the odds of confidence in high level.
- At a strength level of 12%, a 1-unit increase in re-

Table 2: Mixed Effects Logistic Regression Models for Confidence

	Dependent variable:	
	Confidence	
	Full Model M3	Additive Model M4
StrengthLevel4.5	0.435 (−0.134, 1.004)	0.265*** (0.108, 0.423)
StrengthLevel8	0.714** (0.157, 1.272)	0.408*** (0.251, 0.565)
StrengthLevel12	0.898*** (0.357, 1.438)	0.804*** (0.647, 0.961)
StrengthLevel25	1.525*** (0.979, 2.072)	1.870*** (1.703, 2.036)
StrengthLevel40	1.696*** (1.124, 2.267)	2.480*** (2.302, 2.658)
ResponseTime	−2.343*** (−2.753, −1.934)	−2.291*** (−2.473, −2.108)
StrengthLevel4.5:ResponseTime	−0.176 (−0.744, 0.392)	
StrengthLevel8:ResponseTime	−0.325 (−0.886, 0.237)	
StrengthLevel12:ResponseTime	−0.103 (−0.650, 0.445)	
StrengthLevel25:ResponseTime	0.386 (−0.179, 0.951)	
StrengthLevel40:ResponseTime	0.925*** (0.315, 1.536)	
Constant	1.397*** (0.674, 2.121)	1.342*** (0.707, 1.977)
Observations	12,000	12,000
Log Likelihood	−5,569.218	−5,580.312
Akaike Inf. Crit.	11,164.440	11,176.620
Bayesian Inf. Crit.	11,260.540	11,235.760
Note:		*p<0.1; **p<0.05; ***p<0.01

sponse time is associated with an estimated 91.34% decrease in the odds of confidence in high level.
- At a strength level of 25%, a 1-unit increase in response time is associated with an estimated 85.87% decrease in the odds of confidence in high level.
- At a strength level of 40%, a 1-unit increase in response time is associated with an estimated 75.79% decrease in the odds of confidence in high level.

Then we compare the odds of confidence in high level under different strength levels under the constant response time as 1.0:

- Holding the response time constant at 1.0, the odds of confidence in high level for strength level 4.5% is estimated to be 1.30 times that of confidence in high level for strength level 1%.
- Holding the response time constant at 1.0, the odds of confidence in high level for strength level 8% is estimated to be 1.48 times that of confidence in high level for strength level 1%.
- Holding the response time constant at 1.0, the odds of confidence in high level for strength level 12% is estimated to be 2.21 times that of confidence in high level for strength level 1%.
- Holding the response time constant at 1.0, the odds of confidence in high level for strength level 25% is estimated to be 6.76 times that of confidence in high level for strength level 1%.
- Holding the response time constant at 1.0, the odds of confidence in high level for strength level 40% is estimated to be 13.75 times that of confidence in high level for strength level 1%.

6.5 Model Prediction

As the threshold defined as 0.5, the test error of our model is about 0.21, so the model M4 overall correctly predicted

the accuracy of TCD 79% of the time. Furthermore, according to the confusion matrix (Figure 9), our model correctly predicted that the confidence will be high on 3304 observations and low on 3354 observations.

Besides, the AUROC (Figure 10) of our model is 0.861, which indicates the excellent performance of M4. The high values of the measurements of sensitivity and specificity also suggest the good prediction ability of M4. The specificity is about 80% and the sensitivity is about 77.6%, so we can correctly predict about 80% low level confidence of all the low level confidence; and we can correctly predict about 77.6% high level confidence of all the high level confidence. We have better results in the confidence model compared with the model for accuracy because the proportion of the high level of confidence is approximately about 50%. Additionally it is possible that the predictors have a stronger statistically relationship with confidence compared to accuracy.

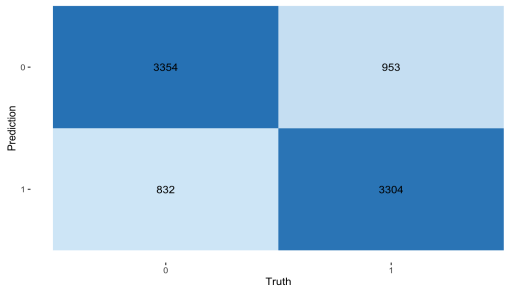

Figure 9: The confusion matrix for the Confidence model

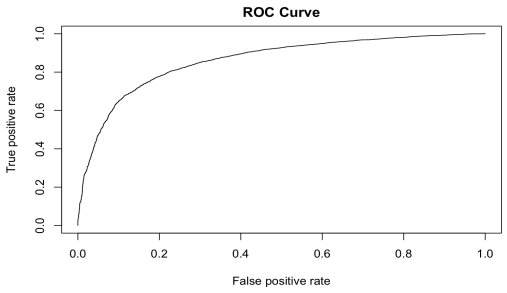

Figure 10: The ROC curve for the Confidence model

7 Conclusion

From the behavior data set, the randomized block model and logistic regression models help to discover the relationship between response time, evidence strength, participants, accuracy and confidence. Stronger evidence leads to faster responses of decisions. As for the accuracy of a TCD, when keeping the evidence strength at a certain level, the increase of the response time causes the decrease of the odds of an accurate decision. As sensory evidence gets stronger, the decrease in percentage of the odds of accurate decision becomes larger. Thus, a slower response is more likely associated with an inaccurate decision. Whereas, when the response time is held constantly, stronger evidence is more likely to link to accurate decision. As for the two levels of confidence, under certain evidence strength, the increase of the response time leads to the decrease of the odds of high level confidence. And when the response time is held constantly, stronger evidence is more likely to lead to higher confidence. Considering the prediction of accuracy and confidence, we have good performance of classification for both variables measured by our test error and AUROC. Indicating that the statistical relationships in the data were significant enough to merit strong predictions. Under the prediction of the accuracy and confidence, we have the good performance of classification with the measurements of test error and AUROC. Overall, our findings support the drift diffusion model and imply the existence of evidence accumulation in decision formation.

7.1 Future Work

Due to the imbalanced accurate data, the sensitivity of the prediction is really small. For future work, we may be interested in collecting more observations with inaccurate TCD or apply a different samplings methods and classification algorithms. Additionally, we may want to take the ordinal property of the data and work with cumulative link models.

References

[1] J. I. Gold and M. N. Shadlen, "The neural basis of decision making.," *Annual Review of Neuroscience*, vol. 30, 2007.

[2] R. Kiani, L. Corthell, and M. Shadlen, "Choice certainty is informed by both evidence and decision time," *Neuron*, vol. 84, pp. 1329–1342, 2014.

[3] W. Dou and J. Samaha, "The neural signature of subjective confidence in perceptual decision making," 2020. Neuromatch Conference 2020.