Predicting the level of accuracy of people in different decision-making situations with the approach of statistics and machine learning

Hanie Jalili

Abstract

Psychologists and researchers in the field of cognitive science use several cognitive science models to investigate the functioning of the human brain in different decision-making situations. But our approach in this article is to use statistical models to predict the level of accuracy of people and their decision time to choose one way from among different ways.

The purpose of this project is to identify the factors affecting the speed and accuracy of people in decision-making, so that after analyzing and examining these factors, we can control them and improve the accuracy of people.

To collect the data for this article, we asked different people two-choice questions and asked them to choose the answer that they think is correct. Then we recorded the collected observations.

To model the data, we used logistic regression and gamma regression statistical models and used machine learning approach to analyze them.

The findings show that the mentioned models were well fitted on the data and have high accuracy for predicting people's accuracy in answering and their answering speed.

Therefore, the results of the model can be used to improve people's accuracy in critical decision-making situations and psychological issues.

Keywords: Logistic regression - Gamma regression– Accuracy in decision making– Decision time– machine learning– Statistical models– Cognitive science - prediction

1) Introduction and statement of the problem

There are many complexities in psychology and psychological issues and in general in the sciences related to the human body. Today, with the spread of artificial intelligence and machine learning models, experts have turned to studying the behavior of the human brain in various situations, and scientists

are trying to make great progress in the field of artificial intelligence by studying the behavior of the brain.
On the other hand, psychologists and cognitive science specialists are also people who need to study the functioning of the human brain in order to analyze human behavior in different situations. In this field, many cognitive science models have emerged and it can be said that the purpose of all of them is to explore human behavior in different situations in life; For example, there are famous psychological models such as the EZ_diffusion model to analyze human behavior in decision-making situations. But our approach is to use statistical models instead of cognitive science models for this issue.
The decision-making process in the human brain is a complex but important process that if we can analyze it using statistical modeling, we can predict the accuracy of people in different decision-making situations and use it to improve the quality and validation of people's decisions. We can also carefully study the behavior of humans in critical situations, which is very useful in psychological science and will increase the quality of life of humans and their mental health.

2) Research background

A famous psychological model for analyzing human behavior in the decision-making process is the EZ_diffusion model. The authors of this model put people in different tests and asked them double-choice questions and asked them to choose the correct answer according to the form of the question. Then they measured their response time and their accuracy in answering the questions (accuracy in this model is equal to the ratio of the number of correct answers to the total number of questions for each person) and as input in the model, the average response time, the variance of the response time and the ratio of the number They put the correct answers. Because these parameters are influential in people's decision-making and can also be achieved and quantified through a scientific experiment.
Psychologists believe that in addition to these parameters, other criteria are also influential in the human decision-making process, which cannot be measured directly. The purpose of presenting this cognitive science model is to study and access these hidden criteria. These criteria are: the level of conservatism of the respondent, the level of difficulty of the question and the time of non-decision:
A) Respondent's level of conservatism: It shows how much people think about answering the question and need previous information to reach the answer. For example, it can be said that older people are more conservative, and as a result,

their response time will be longer and their accuracy in answering the question will be higher.
b) Difficulty of the question: It is obvious that the more difficult the question is, the more time people need to get the correct answer, and also their error rate in getting the correct answer is higher.
C) Non-decision time: Another important factor is the time when the brain has not yet issued a command to answer the question and is still in the analysis stage of the question and different options. This stage is called decode, which is a step before making a decision to choose the correct option. For example, it can be said that elderly people or people with mental problems such as autism will have more time to not make a decision. In fact, it means that they need more time to understand and digest the question and options.
The EZ_diffusion model provides the above criteria as model output. This allows researchers to study the behavior of the human brain in more detail and to compare people using these criteria. This model is not robust against outlier data. In fact, it means that it is affected by outlier data in the model, and this problem makes the output of the model deviate from the reality of the subject. [1]

3) Introduction of data

To collect data, we asked 19 people two-choice questions and asked them to choose the correct answer according to the form of the question and also to focus on one of the following in answering each question:

- Accuracy: It means to pay more attention to accuracy in answering than speed. (instruction = accuracy)
- Speed: It means to pay more attention to speed of response than accuracy. (instruction = speed)
- Balance: Striking a balance between focusing on speed and accuracy. (instruction = neutral)

The number of collected samples is equal to 15819. The columns in the dataset are:

- Their speed to respond = rt
- Respondent ID = subj
- The correctness or incorrectness of the answers to the questions = correct
- Their thinking position when answering the question (explained in the previous section) = instruction

4) exploratory and descriptive analysis of data

Data visualization helps us to have a good understanding of the behavior of the variables and examine the relationship between them before modeling.

- When respondents focused on accuracy in answering, the distribution of response time is as follows. Most people answered questions between 0.6-0.4 seconds. This is the Chule chart and it is better to use gamma regression to analyze it.

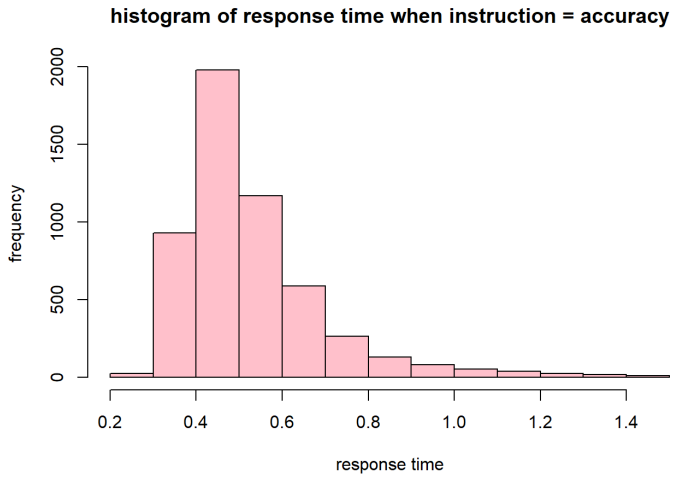

Image number 1: rt distribution when instruction = accuracy.

- Each person has an ID. In this graph, we drew the number of correct and incorrect answers for each of the 19 people separately. Here the number 1 means the right answer and the number 0 means the wrong answer. It can be seen that all people have more number of correct answers.

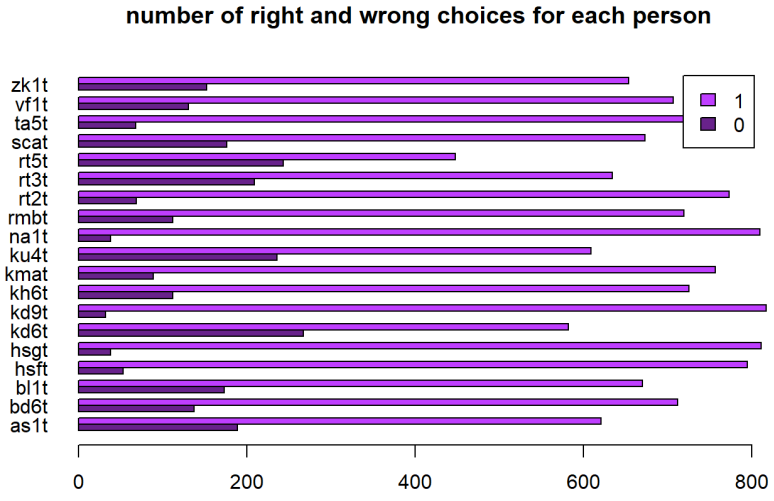

Image number. 2: Chart of correct and incorrect answers of respondents (subj)

- In the graph below, we checked the number of correct and incorrect answers for the instruction variable. Here the number 1 means the right answer and the number 0 means the wrong answer. The number of incorrect answers is higher when people focus on speed than in the other two graphs. This is obvious because when people think about their response speed, they unconsciously make more mistakes. The other two graphs are almost similar to each other in terms of the number of correct and incorrect questions.

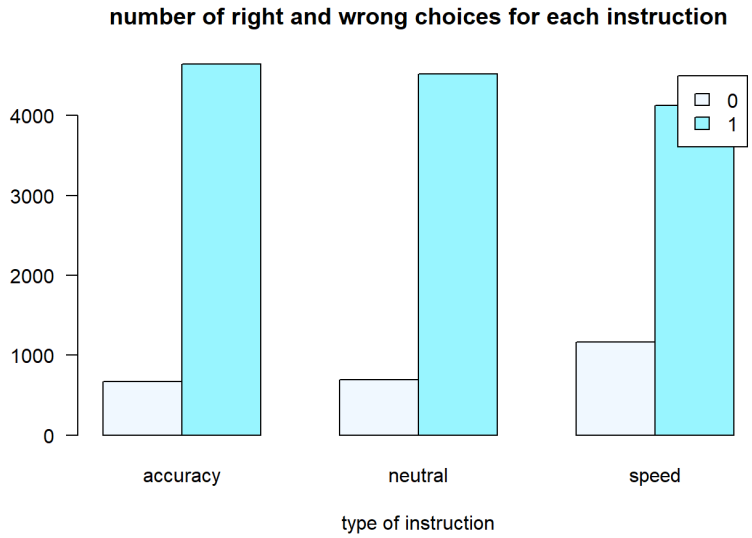

Image number 3: Number of correct and incorrect answers for the instruction variable

- The graph below shows the distribution of people's response time for correct and incorrect questions separately. The maximum response time when people answered questions incorrectly was less than 0.4 seconds. The maximum response time when people answered the questions correctly is more than 0.4 seconds. That is, when people have answered earlier, the probability of their answer being wrong is higher.

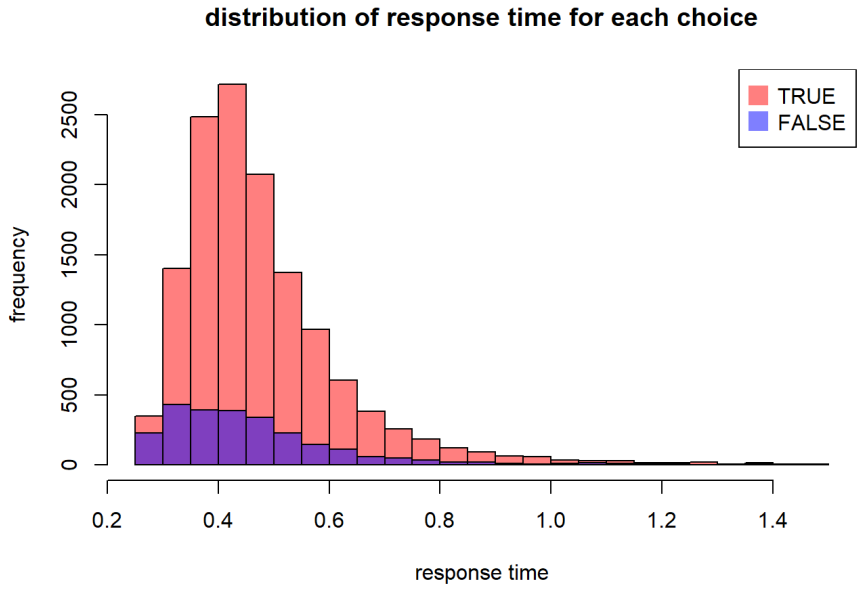

Image number 4: Distribution of the rt variable for correct and incorrect answers

- The following graph is related to the response rate of people for the instruction variable. It is quite evident that when people focused on speed in answering, they answered the questions in less time and when they focused on accuracy, they allocated more time to answer the questions. When they had a balanced behavior, the response rate is the average of the other two graphs.

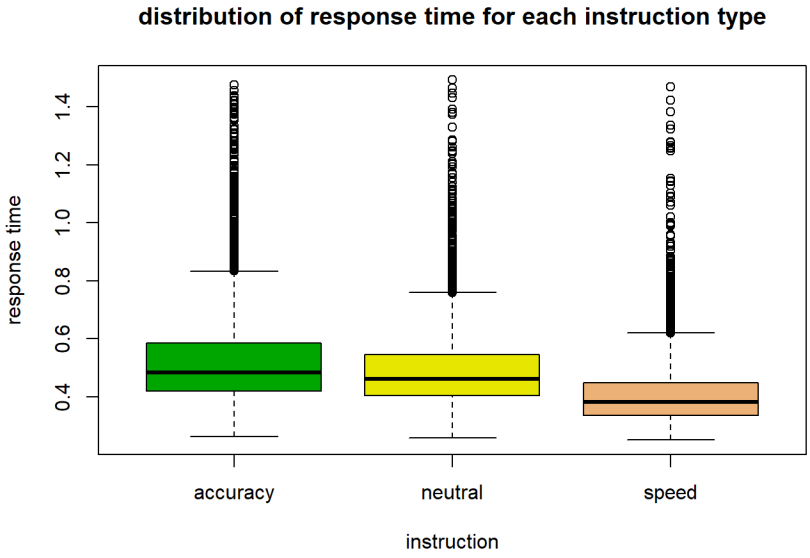

Image number 5: Distribution of the rt variable against the instruction variable

- The diagram below is the same as the diagram above drawn in a different way.

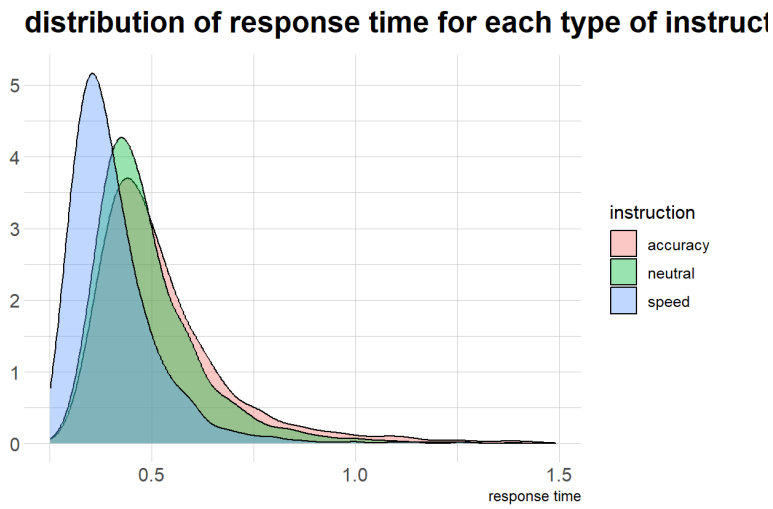

Image number 6: Distribution of the rt variable against the instruction variable

- In the diagram below, we have drawn the 3 variables correct, rt and instruction together; Here, the color represents the response time variable (rt). It can be seen that the response time is less when people focus on speed, and the highest response time and the lowest number of wrong answers happen when people focus on accuracy.

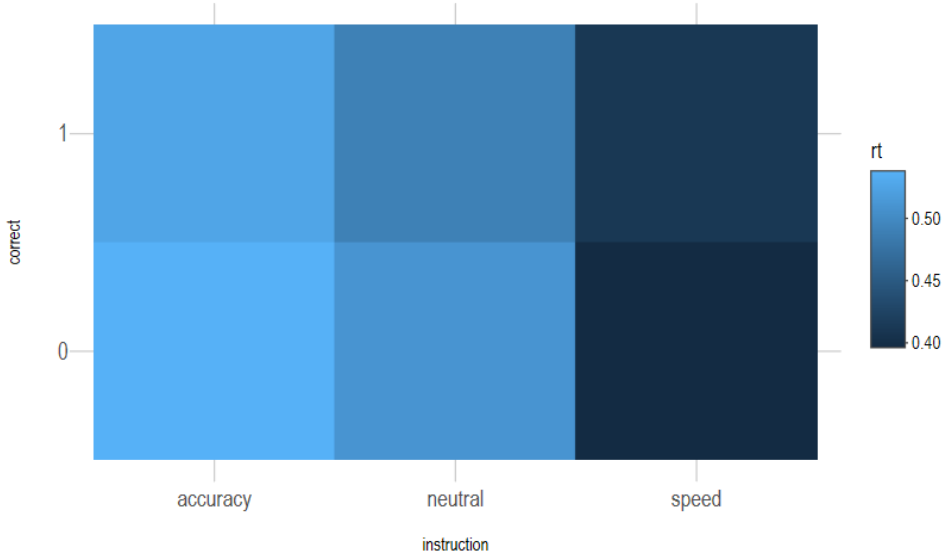

Image number 7: Diagram of correct, rt and instruction variables

Using chi square test to find the relationship between variables:

In order to understand the relationship between the variables, we have used statistical charts so far; Now, for more certainty, we use the chi-square test. In

this test, the null hypothesis is that there is no relationship between the variables. At a significance level of 0.05, the results obtained are as follows:

- variable correct and instruction

P_value in this test is less than 2.2e-16; That is, the null hypothesis is rejected and this shows that there is a significant relationship between the variables.

- variable correct and subj

P_value in this test is less than 2.2e-16; That is, the null hypothesis is rejected and this shows that there is a significant relationship between the mentioned variables.

- variable instruction and subj

P_value in this test is equal to 0.01763; That is, the null hypothesis is rejected and this shows that there is a significant relationship between the mentioned variables.

5) Modeling

In this project, we applied two regression models on the mentioned data:

A) logistic regression model:

Logistic regression model is a statistical model used to analyze the relationship between a binary response variable and one or more explanatory variables. It is a type of generalized linear model (GLM) that assumes a binomial distribution for the response variable and uses a logit link function to model the probability of success. In this article, the answer variable is the accuracy of people in answering the questions (correctly) and we considered the rest of the columns in the data set as explanatory variables.

In the logistic regression model, we are interested in modeling the probability of success of the binary response variable given the values of the explanatory variables. For example, Y = 1)). This probability is expressed as P(Y = 1) orPi(x), where x represents the values of the explanatory variables.

The logit function is used to convert the probability of success into a linear form. logit The probability of success is defined as the logarithm of chance, which is the ratio of the probability of success to the probability of failure:

$$logit[\pi(x)] = log[\frac{\pi(x)}{1-\pi(x)}]$$

The logistic regression model assumes that the logit of the probability of success can be expressed as a linear combination of explanatory variables:

$$logit[\pi(x)] = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$$

Here alpha indicates the intercept or width from the origin of the model. b1، b2، ...، b□ Coefficients associated with explanatory variables x1, x2, ..., x□, and p is the number of explanatory variables.

To get the probability of successPi(x) from logit, we use the inverse of the logit function, which is a logistic function:

$$\pi(x) = \frac{It\ is^{\alpha + \beta 1x1 + \beta 2x2 + \ldots + \beta \square x \square}}{1 + It\ is^{\alpha + \beta 1x1 + \beta 2x2 + \ldots + \beta \square x \square}}$$

Coefficient beta In the logistic regression model, the increase or decrease of the S-shaped curve for the probability of successPi(x) determines One beta Positive indicates that the probability of success increases with increasing values of the corresponding explanatory variable, while beta Negative indicates a decrease. Greatness beta Determines the slope of the curve. [2]

B) Gamma regression model:

Gamma regression is a statistical modeling technique used to analyze positive random variables. It assumes that the dependent variable follows a gamma distribution and its mean is related to a set of explanatory variables through a linear predictor with unknown coefficients and a link function. The link function can be identity, inverse or logit function. In this article, the response variable is the speed of people in responding (rt) and we considered the variables "correct" and "instruction" as explanatory variables.

Gamma regression models can also include a shape parameter that may be constant or dependent on the explanatory variables via a link function, usually the logarithm function. The shape parameter determines the shape of the gamma distribution and can account for the heterogeneity in the data. In gamma regression models with constant shape parameter, the average regression structure is defined as follows.

$$g(\mu_i) = \eta_i = x_i^T \beta$$

Here g is the bond function,b The vector of mean regression parameters, x_i The value of the vector i is the explanatory variables, andthe_i It is a linear projection. The link function, g, is the relationship between the meansm_i and linear predictionthe_i determines [3]

6) Results

After applying the aforementioned regressions, it is necessary to measure the accuracy of these models for predicting new data with a machine learning approach. For this, we need to divide the data into three parts: training data, validation data, and test data. Then we train the models on the training data and finally apply the predicted model on the validation data to measure the accuracy of the model.

We also used criteria such as AIC and various statistical charts such as ROC to select the best model and measure their accuracy.

A) Logistic regression results

First, we train a logistic regression on the training data and then calculate its accuracy on the validation data. In the following, we use selection methods such as backward and forward to select the best available model.

The findings show that all the variables in the model are significant and the accuracy in all these models is the same and almost equal to 83%. While the accuracy on the test data is almost equal to 84%, which shows that our model can predict the new data very well.

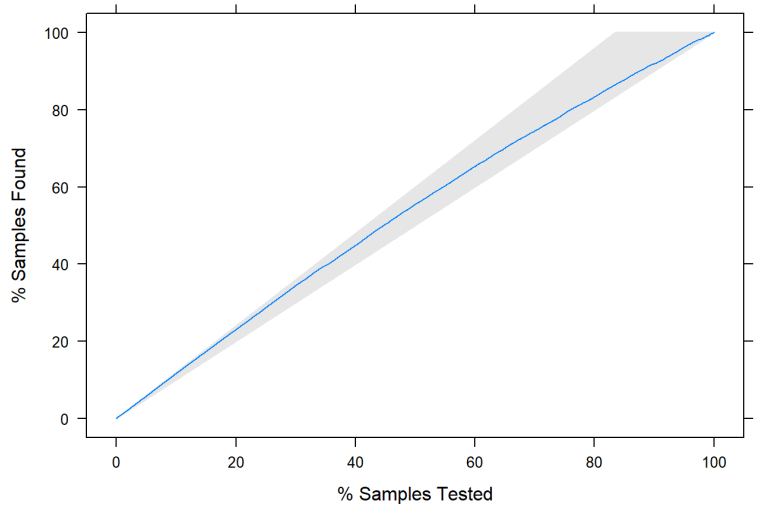

Image number 8: lift chart for logistics model

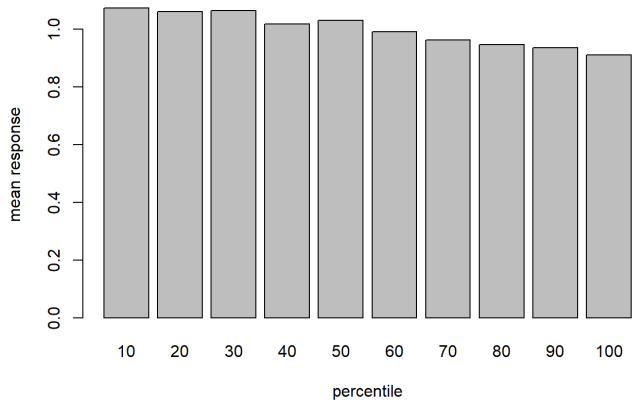

Image number 9: percentile lift chart for logistic model

The charts above as well as the roc chart below also show the adequacy and acceptability of the logistic regression model. (The roc value is also 0.71, which is a significant value.)

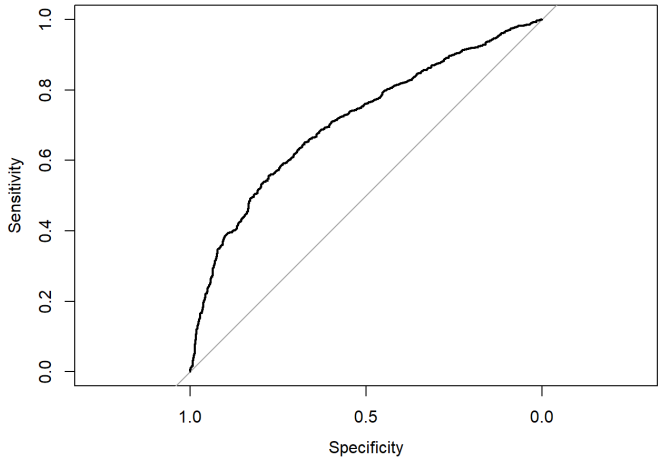

Image number 10: Roc chart for the logistic model

B) Gamma regression results

By applying gamma regression with the identity link function on the validation data, the accuracy of the model is equal to 63%. While the accuracy of the model on the training data is equal to 65%. This shows that our model has very good prediction accuracy.

To better understand this issue, we plotted the actual distribution of the response time variable of people; Also, by obtaining shape and scale parameters from the output of the gamma regression model (using the following formula), we also drew the predicted distribution of people's response time variable. (red curve)

$$shape = \frac{1}{dispersion\ parameter}$$

$$scale = \frac{mean\ of\ gamma\ model}{shape}$$

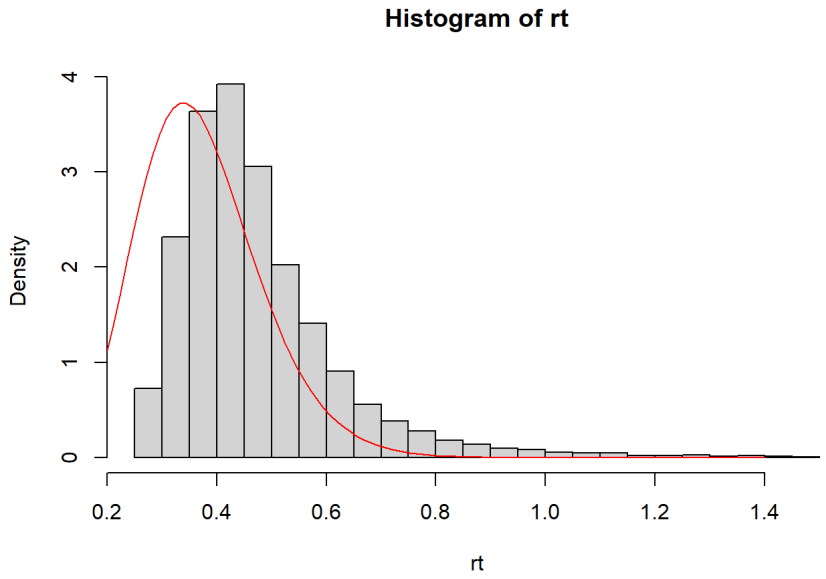

Image number 11: The actual distribution of the rt variable and the predicted distribution of the rt variable

By comparing these two graphs, it can be concluded that the predicted model is highly accurate and is in good agreement with the actual distribution of the response time variable of people (rt).
Finally, for more certainty, we simulated 1000 samples of the gamma regression model (green curve) and compared it with the real distribution of the rt variable. The result is consistent with the previous results and indicates the appropriate accuracy of the model.

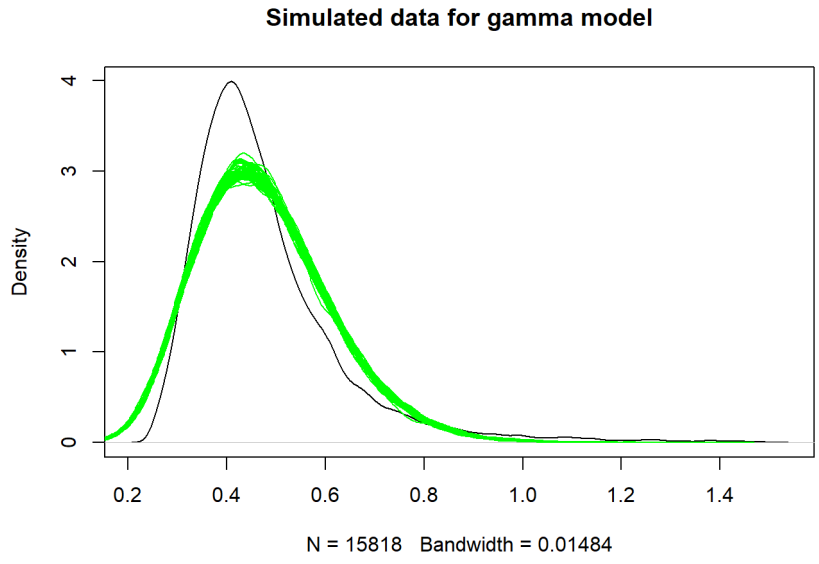

Image number 12: The real distribution of the rt variable and the simulated distribution of the rt variable

7) Discussion and conclusion

The presented models correspond well to reality. Therefore, they can be used to predict people's behavior and guide them to make the right decisions in different situations.

To improve the presented models, we can use more variables when collecting data and ask people to provide us with more information.

8) Resources

- Wagenmakers, E.-J., Van Der Maas, H. L. J., & Grasman, R. P. P. P. (2007). An EZ-diffusion model for response time and accuracy. Psychonomic Bulletin & Review, 14(1), 3–22. https://doi.org/10.3758/bf03194023
- Agresti, A. (2019). An introduction to categorical data analysis. Wiley.
- Bossio, M.C., & Cuervo, E.C. (2015). Gamma regression models with the Gammareg R package. In-State Communicationsinstica, 8(2), 211. https://doi.org/10.15332/s2027-3355.2015.0002.05