Estimating Whether Peoples' Typing Dynamics Change Over Time

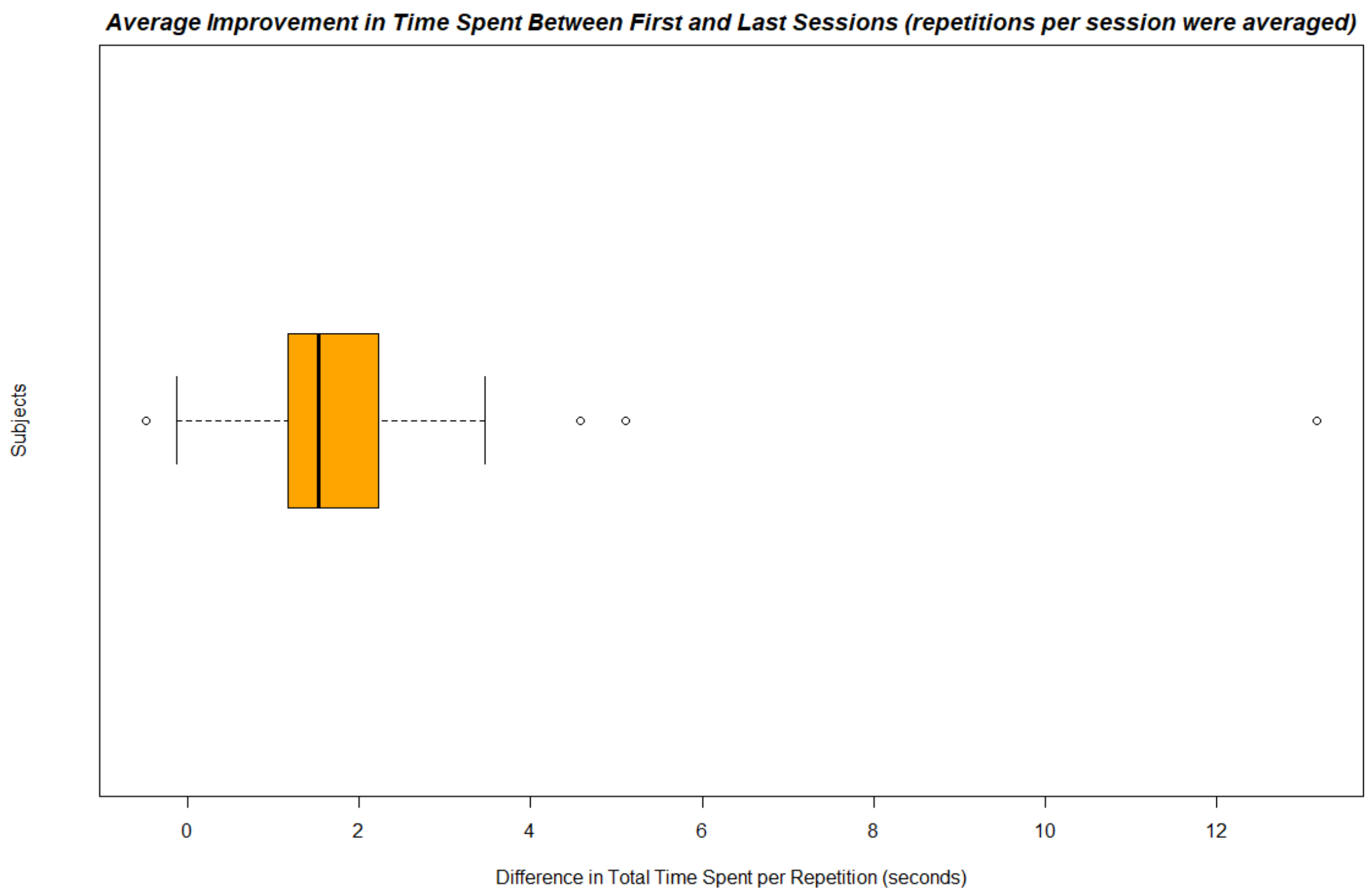
- Analysis by Gavin Gunawardena
- Comparing Anomaly-Detection Algorithms for Keystroke Dynamics by Carnegie Melon University School of Computer Science https://www.cs.cmu.edu/~keystroke/KillourhyMaxion09.pdf

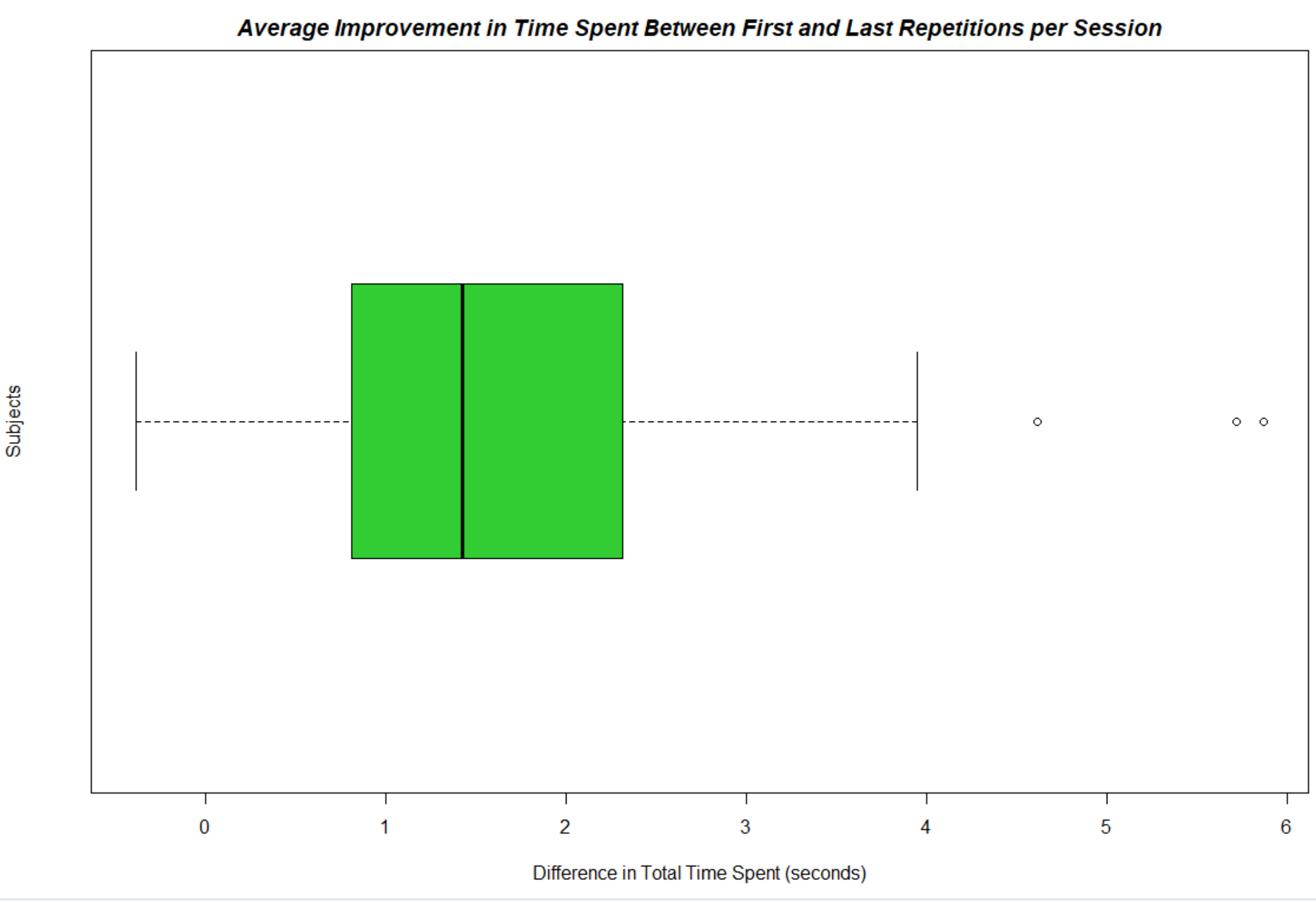
Objective

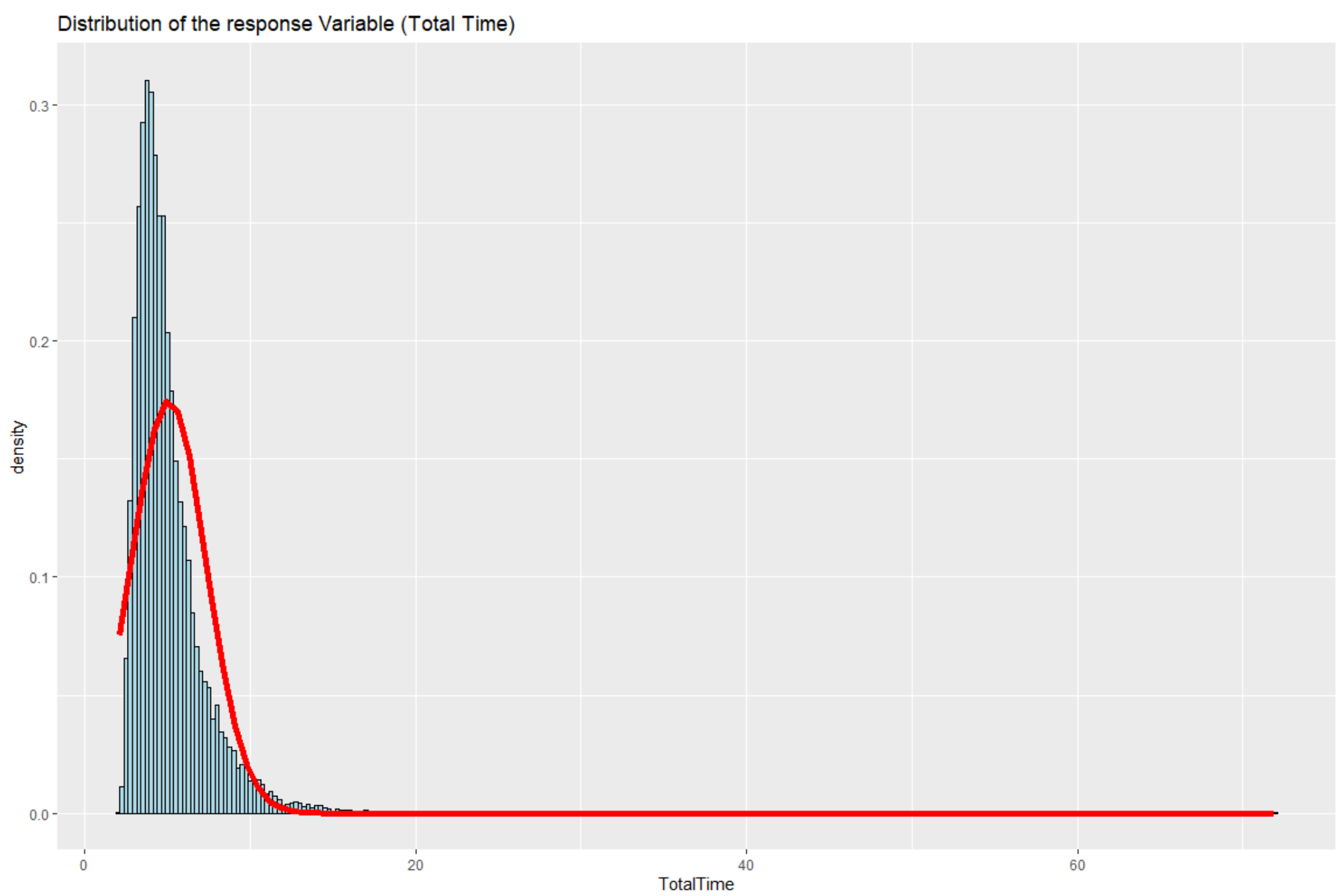
- Attempt to reject the null hypothesis
  - Null Hypothesis: A person's typing dynamics change over time, short term and long term.
- The dataset includes:
  - 51 subjects attempting to type the password: \(.tie5Roanl\)
  - Each subject must type the password 50 times per day, consecutively
  - This is done in 8 consecutive sessions occurring a day from each other
  - 31 fields for actions required to type the password and the time that the subject takes to complete the action
- Response Variable:
  - Time taken to type the password
- Assumptions on dataset
  - Cleaned
  - Verified of any typos and entry errors

Exploratory Analysis Part 1

- Here I manipulated the dataset in order to create some visualizations of it and obtain insights of what I might expect from my statistical analysis. These were used to get an idea on:
  - Average improvement time on password typing speed between first and last sessions
  - Average improvement time on password typing speed between first and last repetitions within a session
  - The distribution of the response variable
- In order to conduct further analysis, I combined the values of the action columns for typing the password to create the column: TotalTime, which will be the response variable


Average Improvement in Time Spent Between First and Last Sessions (repetitions per session were averaged)
Subjects
0
2
4
6
8
10
12
Difference in Total Time Spent per Repetition (seconds)


Average Improvement in Time Spent Between First and Last Repetitions per Session
Subjects
0
1
2
3
4
5
6
Difference in Total Time Spent (seconds)


Distribution of the response Variable (Total Time)
density
0.3
0.2
0.1
0.0
0
20
40
60
TotalTime

Exploratory Analysis Results Part 1

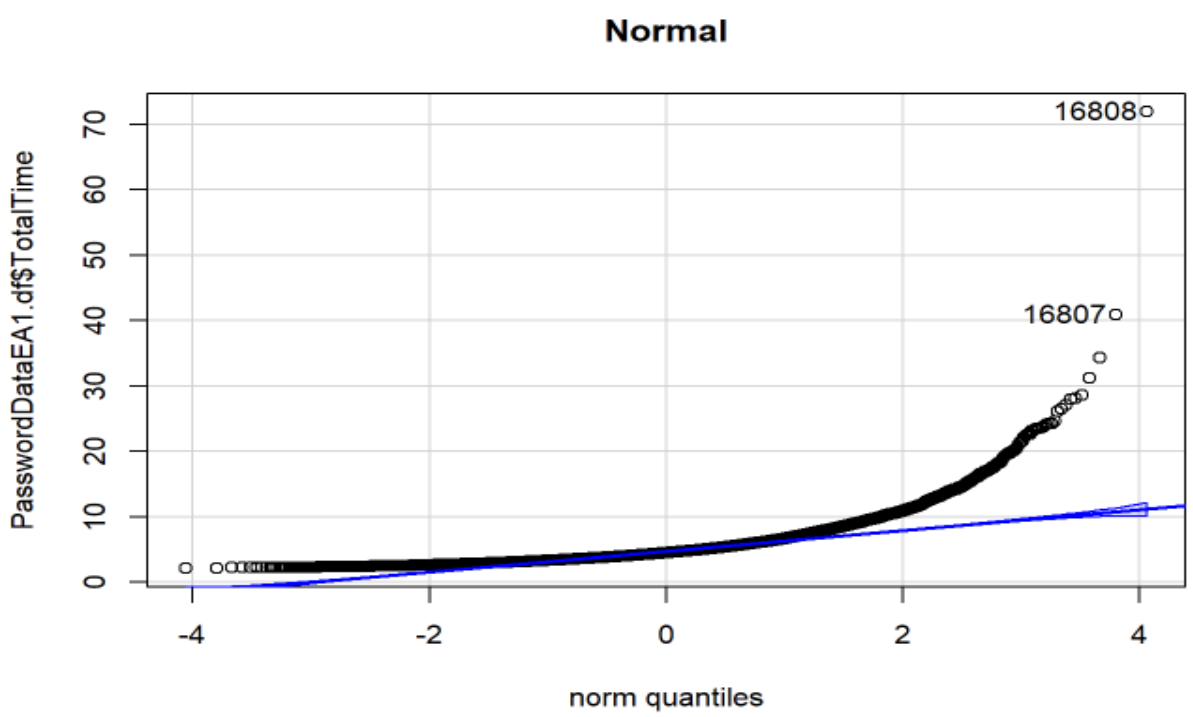
- People do seem to improve in password typing speed as they repeatedly type the password
  - This seems to be more prominent between sessions than between repetitions
- The response variable, total time to type the password, is not a normal distribution and is heavily skewed to the right
  - To preserve the dataset as to not accidentally tamper with results, for this project I've decided to utilize Mixed Effects Models with Penalized Quasi-likelihood to model the data due to its flexibility with abnormal datasets
    - Otherwise I would have to transform the response variable to a normal distribution in order to make use of more common methods such as REML and Maximum Likelihood

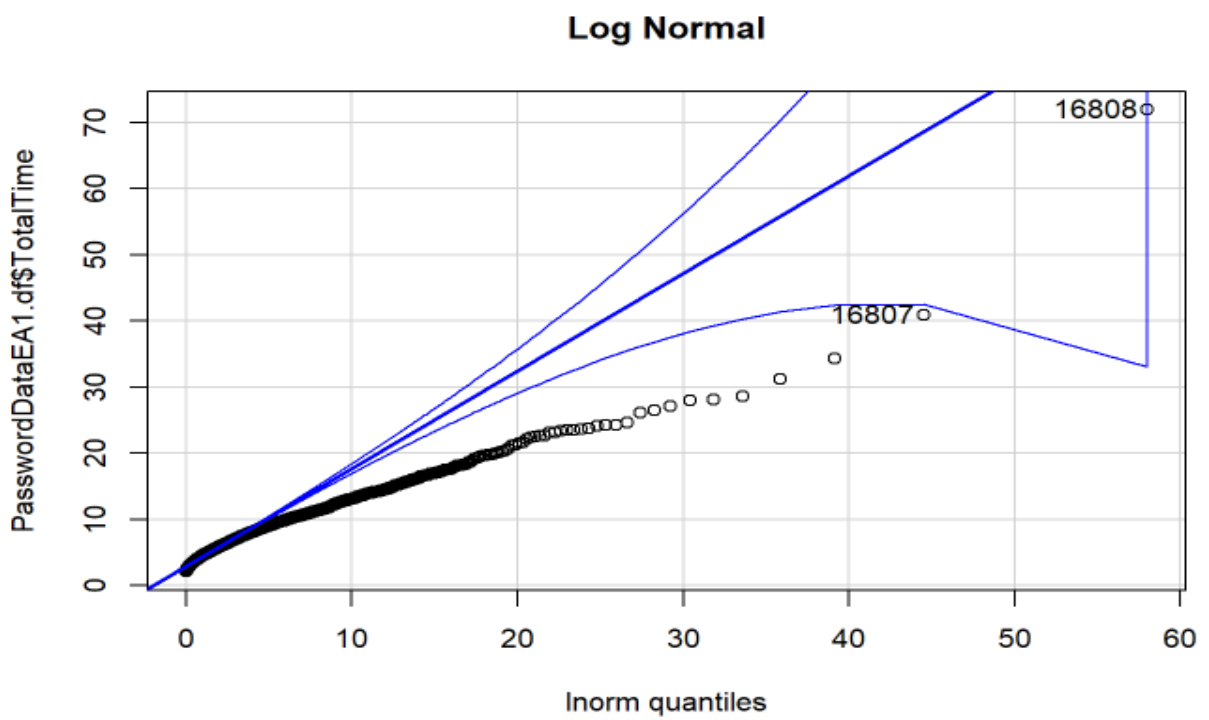
Exploratory Analysis Results Part 1

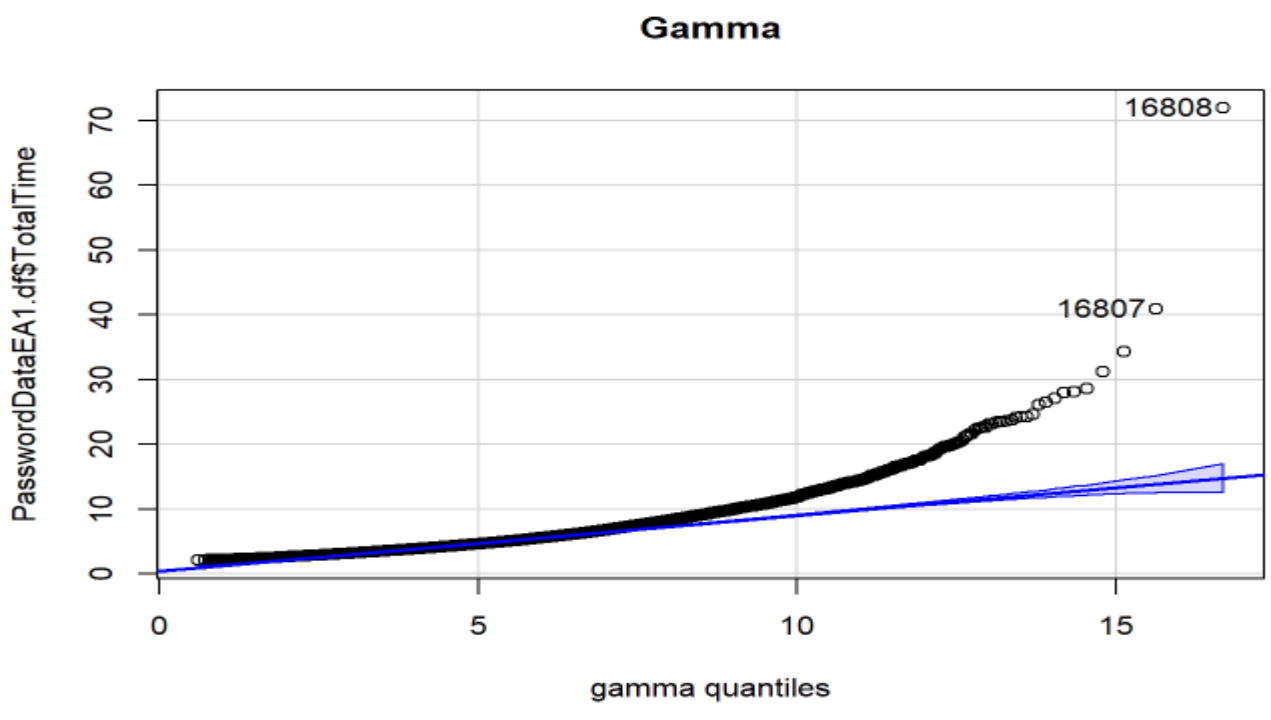
- Models and techniques that will be used
  - Random Intercept Model
    - Commonly used as opposed to generalized linear models for datasets with repeated measurements
    - Assumes that correlation amongst independent repeated measurements on the same unit arises from the shared unobserved variables and that time has a fixed effect
    - Measures dissimilarity in intercepts
  - Random Intercept and Slope Model
    - Similar to random intercept models but measures dissimilarity in intercepts and slopes
  - Penalized Quasilikelihood
    - This technique allows for the fitting of data to mixed effect models when the data is not of a normal distribution.
    - It is an approximate inference technique that allows estimation of model parameters without knowledge of the error distribution of the response variable.

Exploratory Analysis Part 2

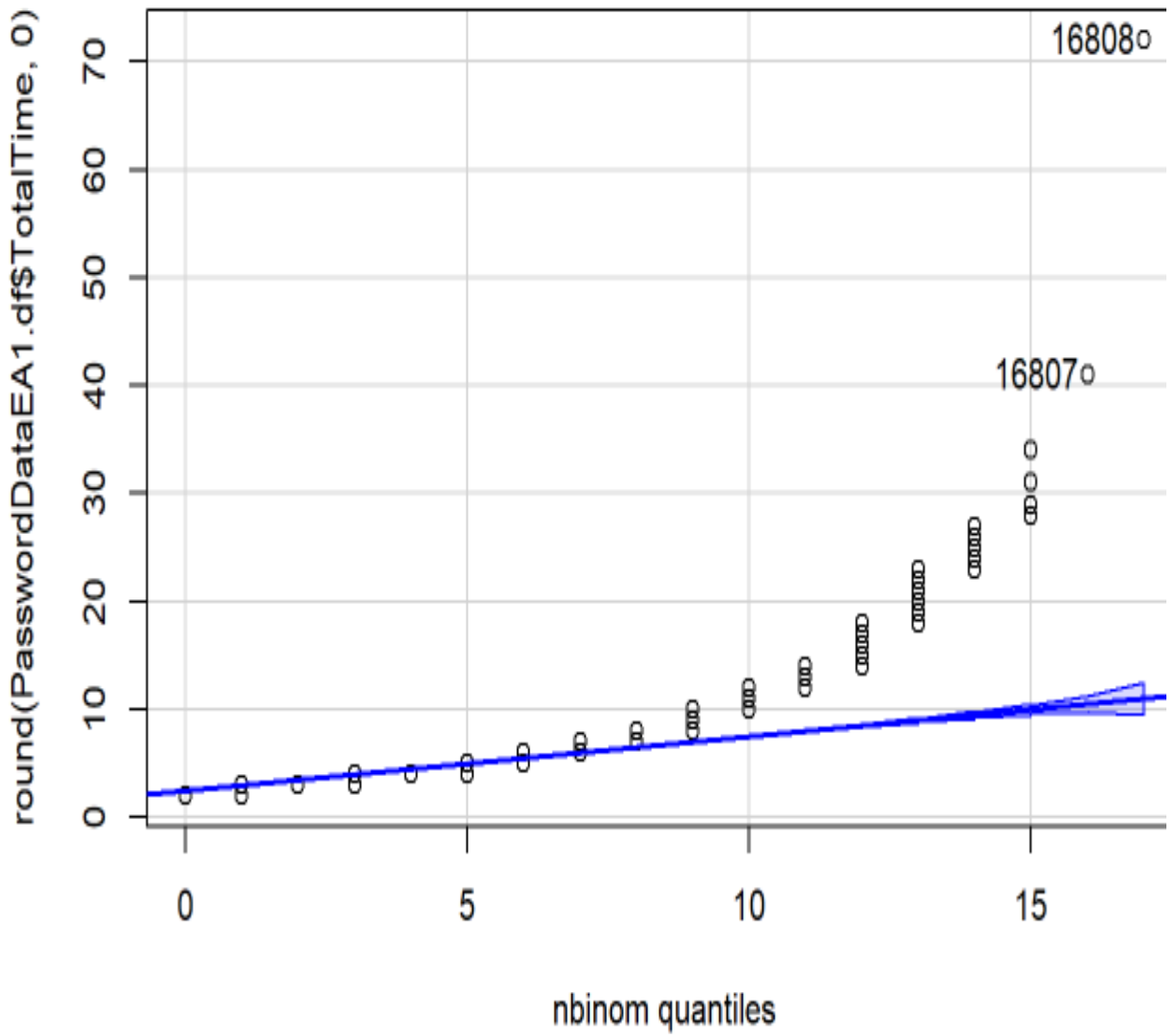
- Here I used visualizations via quantile comparison plots to find what distributions best fit the data, as this will be needed in order to pursue my plan of modeling the data with Penalized Quasi-likelihood
- I Tested between:
  - Normal
  - Log Normal
  - Gamma
  - Negative Binomial
  - Poisson


Normal
PasswordDataEA1.df$TotalTime
norm quantiles
16808
16807
0
10
20
30
40
50
60
70
-4
-2
0
2
4

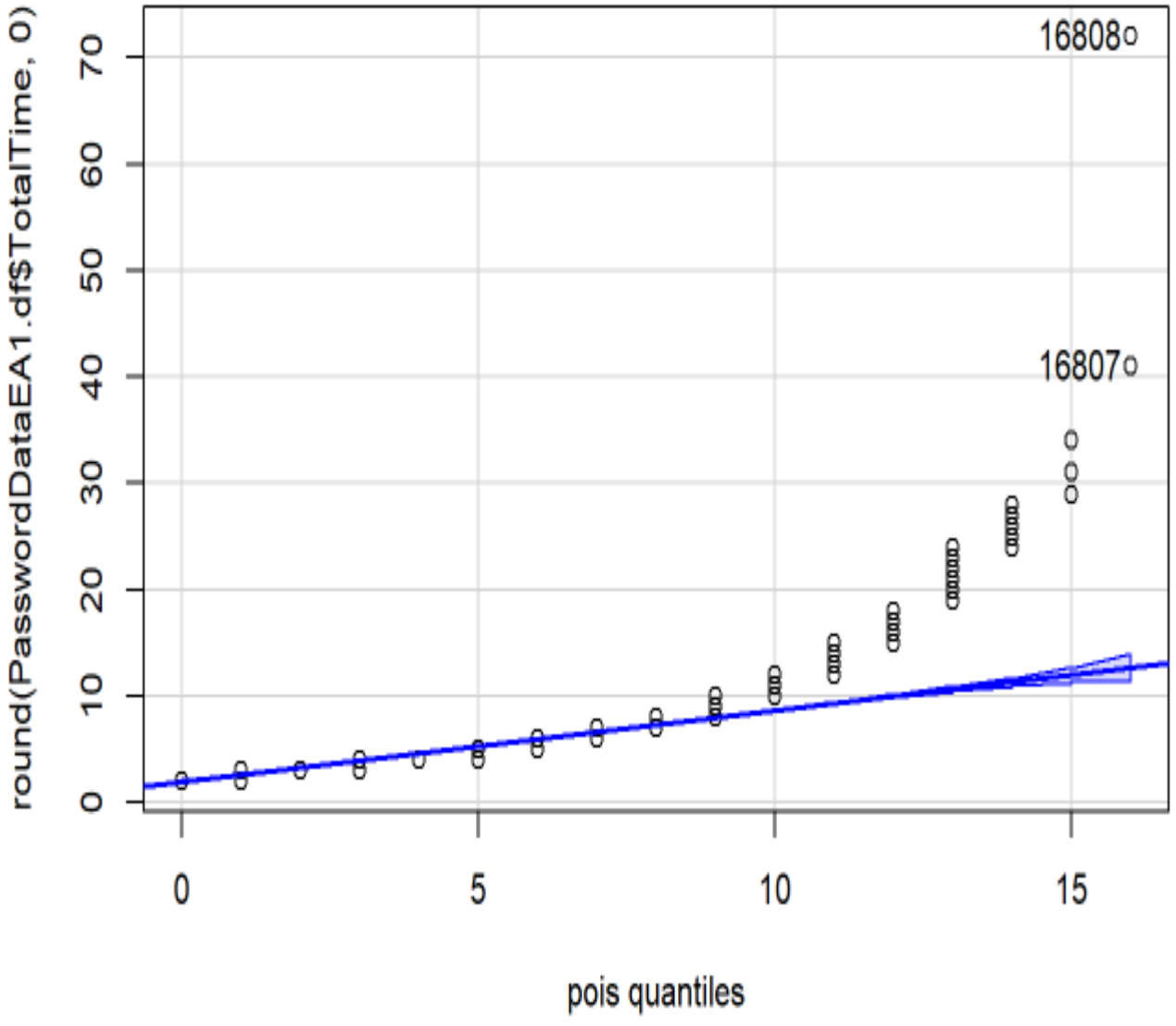

Log Normal
PasswordDataEA1.df$TotalTime
lnorm quantiles
16808
16807
0
10
20
30
40
50
60
70


Gamma
PasswordDataEA1.df$TotalTime
gamma quantiles
16808
16807
0
10
20
30
40
50
60
70
0
5
10
15

Negative Binomial



Poisson



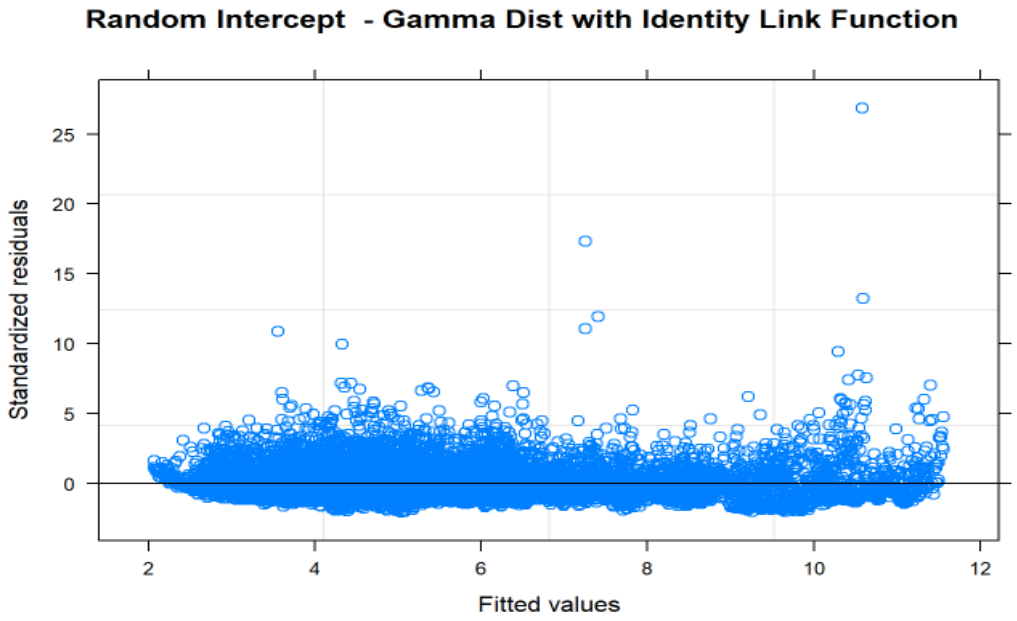
Exploratory Analysis Results Part 2

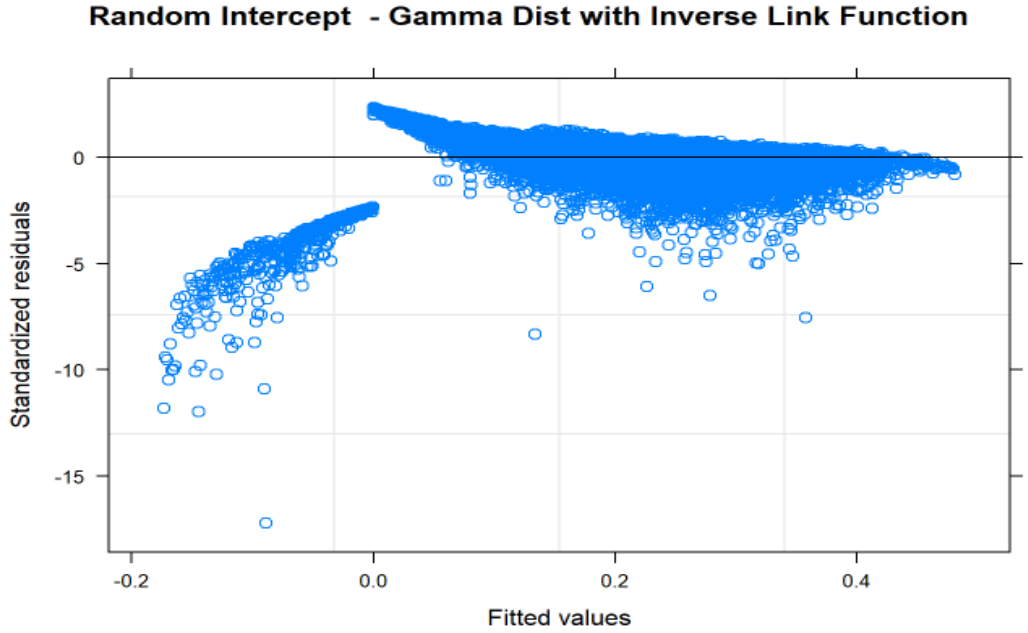
- Negative Binomial, Poisson, and Gamma distributions seem to best fit the response variable
  - I've decided to continue forward with Gamma as although it is comparable to Negative Binomial and Poisson in its data fit, it does not require that the data be discrete values and is compatible with continuous values which the response variable is made up of.

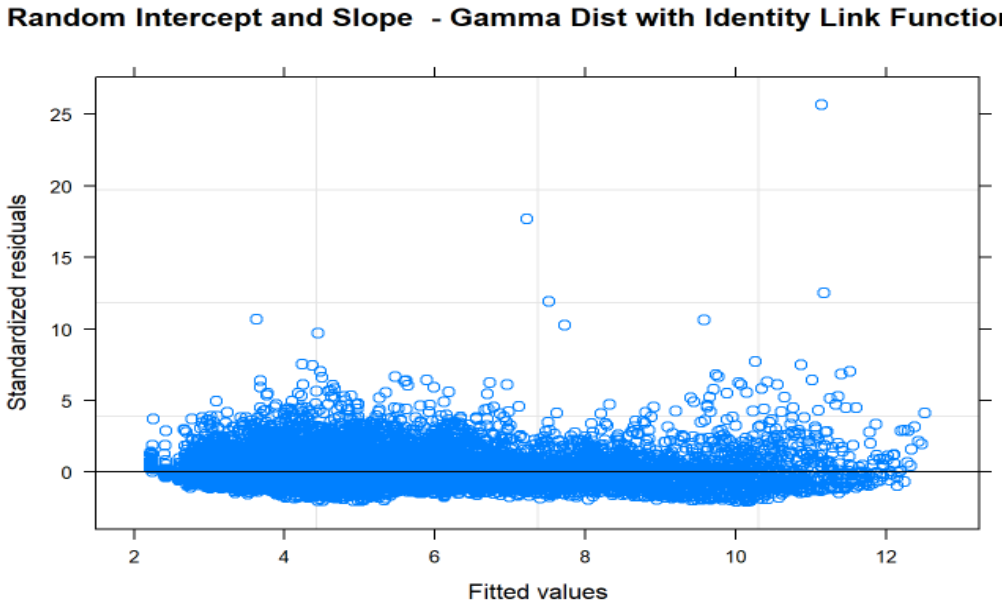
Statistical Analysis

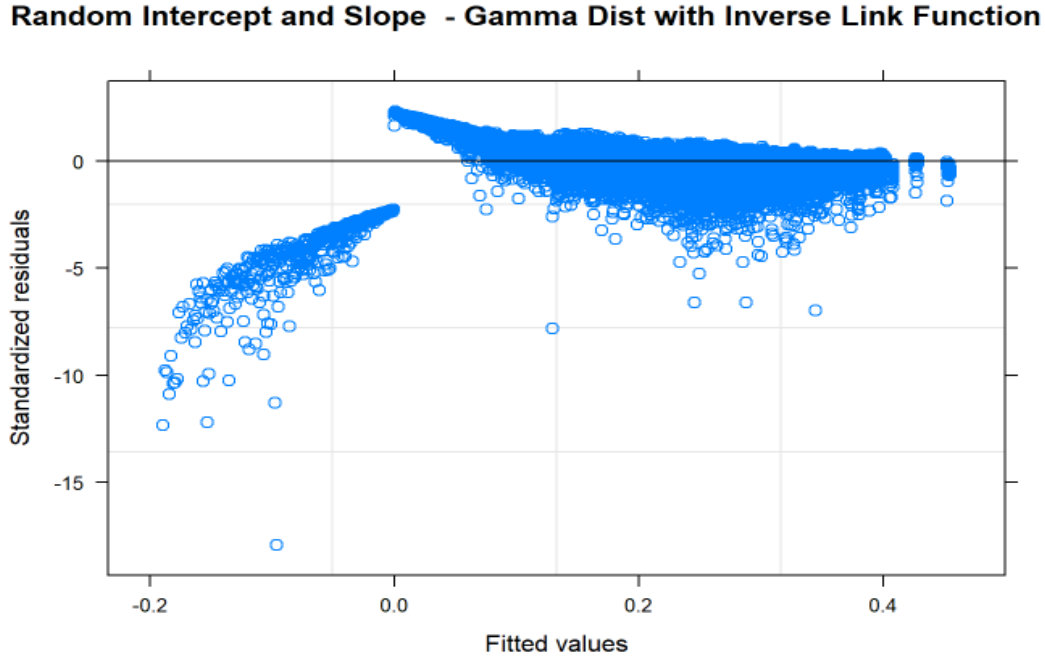
- Here I utilized the glmmPQL function from the LME4 library in R to run a mixed model analysis on the dataset with Penalized Quasi-likelihood
  - Model: Gamma
  - Model Types:
    - Random Intercept
    - Random Intercept and Slope
  - Link functions:
    - Identity
    - Inverse
  - Response Variable:
    - TotalTime – represents total time spent to type the password
  - Fixed effects variables:
    - SessionIndex – represents session number
    - Rep – represents repetition number
  - Random effects variable:
    - Subject – represents the individual person typing the password

Residuals Charts









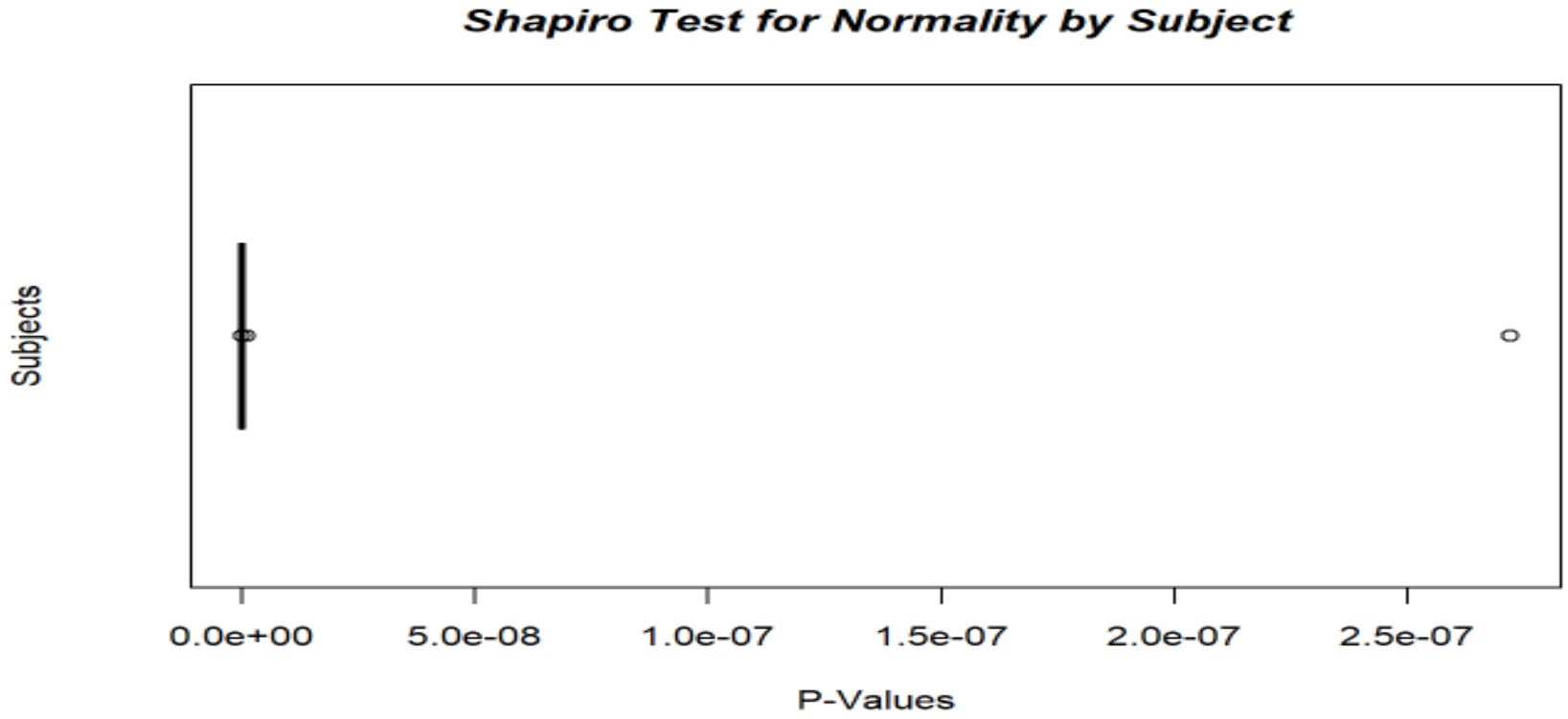
Statistical Analysis Results

	fixed.effect	random.effect	family	fixed.vars	value	std.error	t_value	p_value
1	TotalTime ~ sessionIndex + rep	1 | subject	Gamma(link=identity)	sessionIndex / rep	-0.186639 / -0.007104	0.00293549 / 0.00046170	63.58006 / -15.38636	0 / 0
2	TotalTime ~ sessionIndex + rep	1 | subject	Gamma(link=inverse)	sessionIndex / rep	0.02501722 / 0.00130749	0.000200294 / 0.000011767	124.90233 / 111.11277	0 / 0
3	TotalTime ~ sessionIndex + rep	1+rep|subject	Gamma(link=identity)	sessionIndex / rep	-0.186821 / -0.011385	0.00289161 / 0.00171970	-64.60783 / -6.62033	0 / 0
4	TotalTime ~ sessionIndex + rep	1+rep|subject	Gamma(link=inverse)	sessionIndex / rep	0.02638535 / 0.00050069	0.000222314 / 0.000086832	118.68508 / 5.76625	0 / 0

Note: *Green represents models that best fit the dataset
*Fit was measured via residuals charts shown in the previous slide as glmmPQL does not output AIC, BIC, or Log likelihood values

Post-Hoc Analysis Part 1

- For the post-hoc analysis, I decided to confirm that the data is in fact not a normal distribution via a Shapiro-Wilkes test, as the technique used in the analysis is dependent on the dataset not being normally distributed, as otherwise, different techniques would have been superior.
  - This test has a limit on the number of values in can take in at 3500, so I ran this for each subject and outputted a box plot of the results which showed that none of the subjects had normally distributed total times.:



Post-Hoc Analysis Part 2

- Here I tested for an interaction between the two fixed variables by changing the model to include sessionIndex * rep as a fixed effect as shown here:
  - TotalTime ~ sessionIndex*rep + sessionIndex +rep + (1 | subject)
- Results (below) were that there is an interaction between sessionIndex and rep but it has a very small effect on the response variable compared to that of sessionIndex and rep individually

	fixed.effect	random.effect	family	fixed.vars	value	std.error	t_value	p_value
1	TotalTime ~ sessionIndex*rep + sessionIndex + rep	1 | subject	Gamma(link=identity)	sessionIndex	-0.268579	0.00602308	-44.59160	0
2	TotalTime ~ sessionIndex*rep + sessionIndex + rep	1 | subject	Gamma(link=identity)	rep	-0.022724	0.00109991	-20.66014	0
3	TotalTime ~ sessionIndex*rep + sessionIndex + rep	1 |subject	Gamma(link=identity)	sessionIndex:rep	0.003146	0.00020085	15.66136	0

Conclusion

- Time taken to type a password changes over time as one retypes it throughout the day and consecutive days
  - The change is by about -0.19 seconds per session and -.01 seconds per repetition
- The response variable, total time to type the password, was not normally distributed
- There is an interaction between session and repetition in regard to effect on the response variable, total time, but it is small compared to the individual effects of the variables

References

- Arbor Custom Analytics. (2020, October 27). Mixed models in R: a primer. Retrieved from Arbor Custom Analytics: https://arbor-analytics.com/post/mixed-models-a-primer/
- Everitt, B. S., & Hothorn, T. (2014). A Handbook of Statistical Analyses Using R, 3rd Edition. In B. S. Everitt, & T. Hothorn, A Handbook of Statistical Analyses Using R, 3rd Edition (pp. 139, 247-251). Boca Raton: CRC Press.
- Comparing Anomaly-Detection Algorithms for Keystroke Dynamics. Retrieved from Carnegie Melon University School of Computer Science: https://www.cs.cmu.edu/~keystroke/KillourhyMaxion09.pdf
- Pilowsky, J. (2018, October 19). A Practical Guide to Mixed Models in R. Retrieved from Tufts University Web Site: https://ase.tufts.edu/bugs/guide/assets/mixed_model_guide.html
- Mammen, E., & van de Gee, S. (1997). Penalized Quasi-Likelihood Estimation. *The Annals of Statistics*, 1015. Retrieved from Central University of Finance and Economics: http://lib.cufe.edu.cn/upload_files/other/3_20140520034435_Penalized%20quasi-likelihood%20estimation%20in%20partial%20linear%20models.pdf