Changing Rhythms

A White Paper Offering Insights Into How Keystroke Dynamics Change Over Time

Nikhil Karthik Pamidimukkala

December 10, 2018

Abstract

The efficacy of keystroke dynamics as a biometric authentication system has been an interesting subject of research for many years. In 2009, Dr. Roy Maxion and colleagues collected a keystroke dynamics dataset and compared 14 anomaly detection algortithms [**1**]. Using the keystroke dynamics data they collected, the task is to explore how the typing dynamics of a subject change over time. This paper presents a precise and yet a comprehensive summary of the data analysis performed. A 3 step methodology is incorporated in this analysis to explore the conjecture in question.

Problem Statement

It is quiet common to make assumptions that a person's typing patterns change through practice. However, the statistical significance of this assumption should be established. Using the keystroke dynamics dataset [**1**], a comprehensive data analysis which includes graphical and statistical analysis should be used to explore and validate that a person's typing dynamics change over time.

Background and Data Description

51 subjects were recruited for the keystroke data collection. Each subject participated in 8 sessions which were one day apart. During each session a subject typed the passcode **.tie5Roanl** 50 times [**1**]. For each typing, a set of keystroke features were extracted. The features were:

- **Key-Down-Down(DD.)** :The time between pressing down a key to the time to press down the next key.
- **Key-Up-Down(UD.)** :The time between a key coming up to the time to press down the next key
- **Hold-time** : The amount of time a key is held down.

Approach to the Solution

To address the conjecture in question, a 3-step approach is implemented

- Explanatory Data Analysis
- Perform Appropriate Statistical Analysis
- Testing model Assumptions and Post-Hoc Analysis

Explanatory Data Analysis

The first step to explore the data is to analyze the summary statistics. Analyzing the summary statistics reveal some negative minimum values in UD. features. This is an interesting occurence because negative values in a UD feature indicate that a second key is pressed before the first one is released **[2]**. These negative values can be termed as overlaps when typing the passcode. If an assumptions is made that these overlaps increase over time (sessions), then as consequence the hold times(H features) should increase and latency between holds (UD features) should decrease. To explore this assumption, a bar plot is plotted with the number of overlaps in each session and Loess smoother mapping the change in proportion of overlap occurences which is showin in Figure 1.

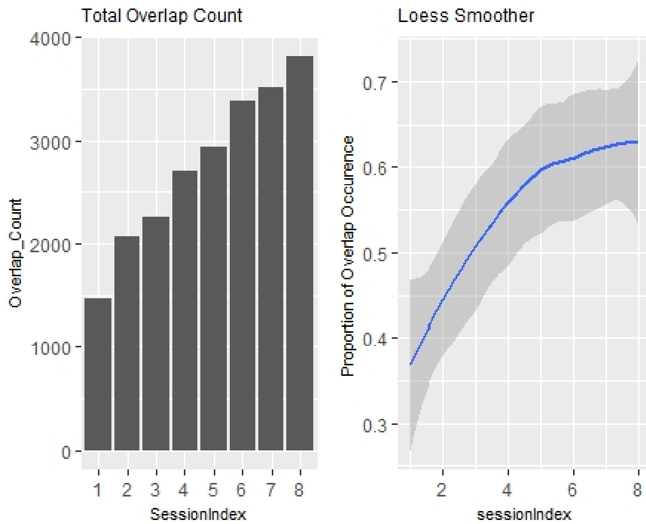

Figure 1: Bar Chart of Overlap Count and Loess Smoother for Proportion of overlap occurrence

The change in hold times, UD times are also explored. Figure 2, shows that Total Hold time (Sum of Hold Features) increase over repetitions and sessions. Total UD time (Sum of UD features) decreases over sessions and repetitions. And total time to type passcode (Sum of DD features and H.Return) decreases over sessions and repetitions.

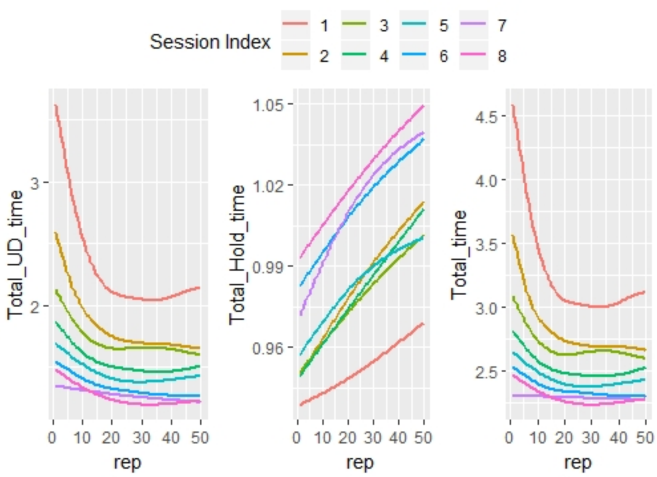

Figure 2: Smoother of Total Hold time, Total UD time, Total time to passcode

Response variable Creation

The Hold times are increasing across sessions and Key-Up-Down times are decreasing across sessions. The pattern is same across repetitions. Therefore, the overlaps increase across sessions and it might be an indication that subjects are expecting the next keys through practice. To infer this using a statistical model, a new response variable is created with some data transformation. Two individual observations i.e. count of occurences of overlaps (Overlap_Yes) and count of non-occurence of overlaps (Overlap_No) for each subject in a session are calculated. Both of these will be combined using a two-column matrix and modelled as the response variable. Since the question of interest how the typing dynamics change over time, change in occurence of overlaps over sessions can be considered as an appropriate choice of response. If the inference through the model that occurence of overlaps increase with sessions is statistically significant, it can be concluded that the conjecture that a person's typing dynamics change over time.

sessionIndex	subject	Overlap_No	Overlap_Yes	Proportion_of_Occurences
1	s002	42	8	0.16
1	s003	27	23	0.46
1	s004	17	33	0.66
1	s005	49	1	0.02

Table 1: First 4 rows of Data for model building

Model Building

A logistic regression model can be considered with the two individual observations i.e. counts of overlap occurrences and counts of non-occurrences grouped as the reponse. However, a Generalized Linear Model (Logistic Regression) makes assumption of independence of observations. The data used for the model is a longitudinal data which have repeated measures from the same subjects thus violating the assumption of independence. A binomial Generalized linear Mixed Effects Model can be used which adds the random effects to account for the non-independence of observations. The primary objective of the analysis is to explore how keystroke dynamics change over time for a subject. Here sessionIndex is the time.Therefore, sessionIndex is chosen as independent variables with subject as the random intercept to account for the subject wise variation. The glmer() function in the lme4 package [**3**] in R [**4**] is used to fit the Generalized Linear Mixed Effects Model.

Fitting a Model

Initially, Generalized Linear Mixed Effects Model with subject as Random Intercept is fit. However, in repeated measures variances of the later repeated measures might be greater than those taken earlier. To account for the observed pattern of covariance between repeated measures , a random slope and intercept model can be considered[**5**]. The better among the two models is determined using a likelihood ratio test.This is done to see whether one model is a significant improvement from other. Performing this test, reveals that the random slope and intercept model is significantly better. Therefore, this model is fitted. The model formula can be written as

```
(Overlap_Yes, Overlap_No) ~ sessionIndex + (sessionIndex|subject)
```

Inference from the fitted model

The coefficients of the fixed effects (sessionIndex) are significant. The positive coefficient of sessionIndex suggests that on average for a subject, the probability of occurences of overlaps increases over sessions. The variance in the random intercept (subject) is high compared to the random slope (sessionIndex). Likelihood Ratio Tests or parametric bootsrapping can be used to further verify significance of predictors. Figure 3 shows the predicted probability of occurences of overlaps increases over sessions. Increase in overlaps suggest the the subject's become familiar with keystrokes and start expecting the next keys through practice.

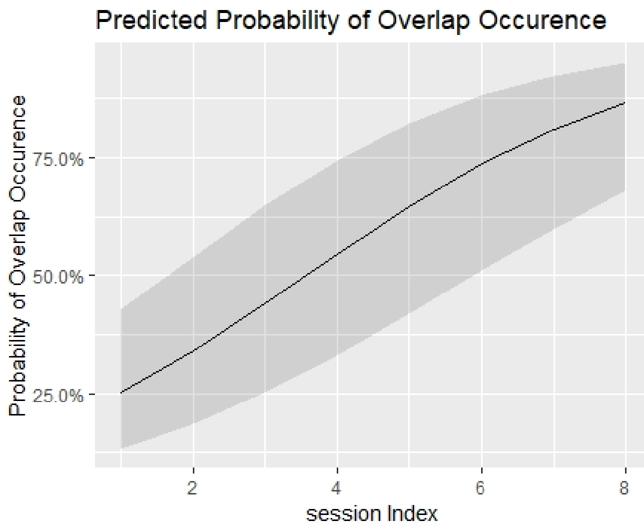

Figure 3: Predicted Probability of overlap occurence

Model Assumptions

To verify that the model is providing trustworthy estimates, it is important analyze model diagnostics. In mixed models, two levels of residuals need to be check. First, normality of random effects needs to be verified. Figure 4 shows the Q-Q plot of the random effects. It can be seen that both the random effects are normal as the Q-Q plot is approximately a straight line.

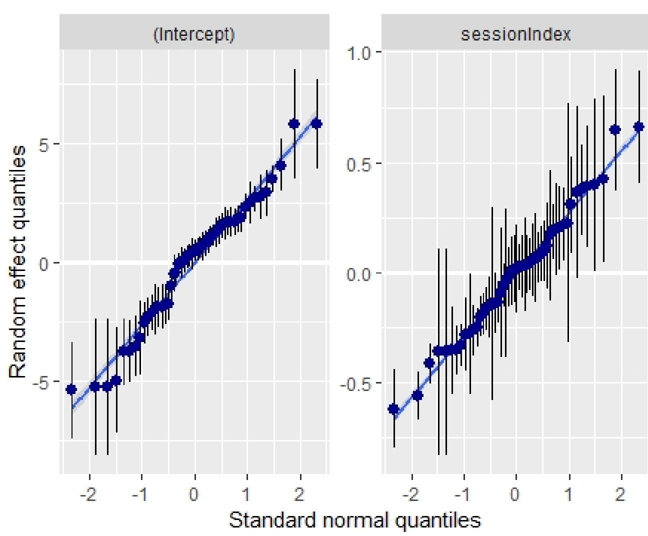

Figure 4: Q-Q plot of random effects

The next step to check the normality of residuals. In Generalized linear mixed models, small deviations from residual normality is tolerable. However, they should be analyzed to see gross violations. Figure 5 shows the residuals plots which include Q-Q plot of residuals and Fitted vs Residual graph. The Q-Q plot shows an approximately straight line but with slight heavy tails. The Fitted vs Residuals graph show homoscedasticity of residuals without showing any patterns. It can be concluded that the model diagnostics are good.

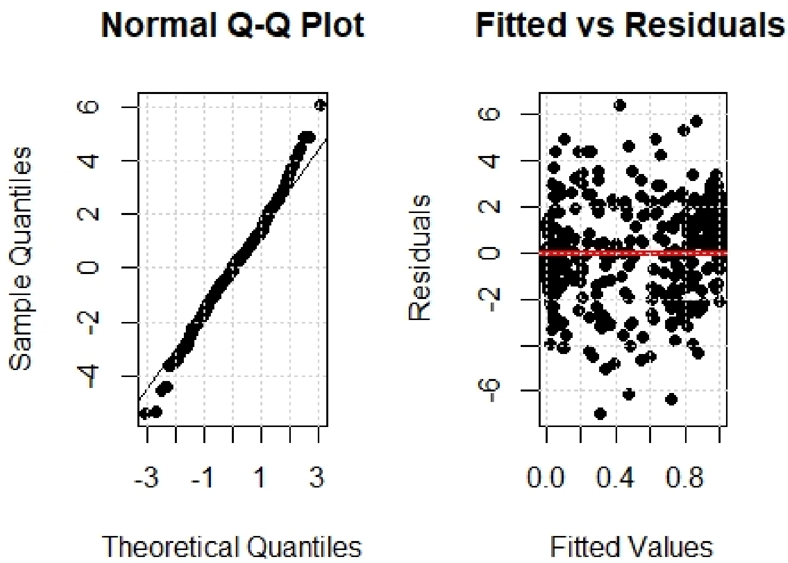

Figure 5: Residual Plots

Post-Hoc Analysis

Post-Hoc Analysis can be done to validate the significance of effects determined in the model along with potential pairwise comparisons. This can be be done using techniques such as bootstraping , likelihood ratio tests etc. A likelihood ratio test is done to test compare significance of the predictor. The Null model (model with only random effects) is compared with the full model. The test reveals that the random intercept and slope model fitted is a significant improvement from the null model and it has the lowest AIC. Therefore, sessionIndex can be considered to have a significant impact on the response. Similarly, the significance of the random effects is tested. A logistic regression model with no random effects is compared with the full model. The random intercept and slope model can be concluded as a significant improvement from the logistic regression model not only because of the significance of likelihood ratio test but also because of largely different AIC values of the models.

Conclusion

The inference that can be drawn from the graphical and statistical analysis done is that typing dynamics for a subject changes over sessions. The conclusion from the statistical model built, is that the probability of occurences of overlaps increases over time (sessions). Increase in overlaps means that the Hold times on keys increase and latency between holds decrease. This indicates that the subjects are able to expect the next key through practice.

References

[1] Killourhy, K. S., & Maxion, R. A. (2009). Comparing anomaly-detection algorithms for keystroke dynamics. 2009 IEEE/IFIP International Conference on Dependable Systems & Networks. doi:10.1109/dsn.2009.5270346

[2] Killourhy, K.S. (2012). A Scientific Understanding of Keystroke Dynamics.

[3] Douglas Bates, Martin Maechler, Ben Bolker, Steve Walker (2015). Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software, 67(1), 1-48. doi:10.18637/jss.v067.i01

[4] R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

[5] Hothorn, T. (2017). Handbook of Statistical Analyses Using R, Third Edition. S.l.: CRC PRESS.

[6] D2l-Stat-601-Module8:Longitudinal Data Analysis and Mixed models-Chapter_13

[7] D2l-Stat-601-Module 2: Logistic Regression and GLM-Ch7

[8] https://strengejacke.wordpress.com/2017/10/23/one-function-to-rule-them-all-visualization-of-regression-models-in-rstats-

[9] https://www.ssc.wisc.edu/sscc/pubs/MM/MM_TestEffects.html#test-of-random-parameters

[10] https://www.ssc.wisc.edu/sscc/pubs/MM/MM_TestEffects.html

[11] Yihui Xie (2018). knitr: A General-Purpose Package for Dynamic Report Generation in R. R package version 1.20.

[12] Letaw Alathea (2015). captioner: Numbers Figures and Creates Simple Captions. R package version 2.2.3. https://CRAN.R-project.org/package=captioner

[13] https://datascienceplus.com/linear-mixed-model-workflow/