Stat 601 Final Project Presentation

Fall 2018

Nikhil Karthik Pamidimukkala

December 10, 2018

Contents

Background

Dr. Roy Maxian and colleagues recruited 51 subjects at CMU. The process of data collection included 51 Subjects typing the passcode **.tie5Roanl** 50 times in an individual session. There were 8 session altogether which were one day apart. During these sessions, several keystroke features have been recorded such as **Key-Down-Down** (The time between pressing down a key to the time to press down the next key.), **Key-Up-Down** (The time between a key coming up to the time to press down the next key) and **Hold-time** (The amount of time a key is held down.). [1]

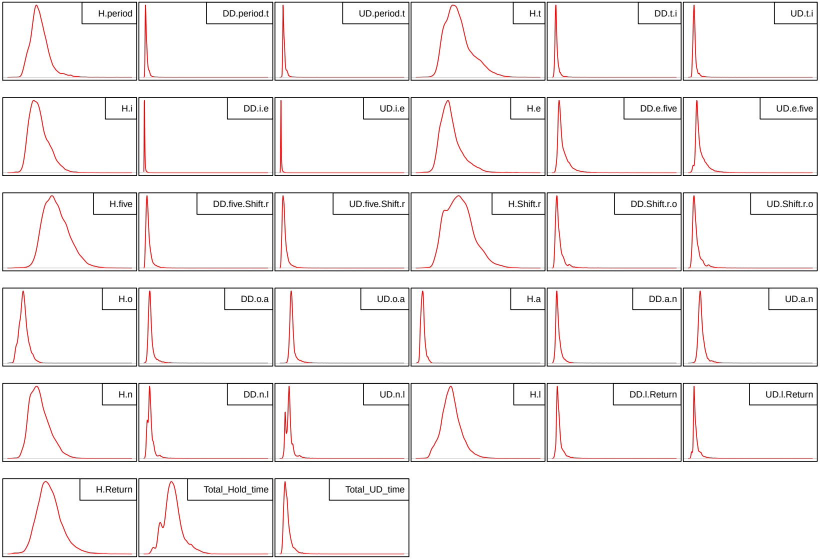
Purpose

The purpose of the project is to explore the conjecture that a person's typing dynamics change over time. To achieve this, the following methodology is used.

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

Exploratory data analysis

- Exploring distribution of each keystroke feature



Exploratory data analysis

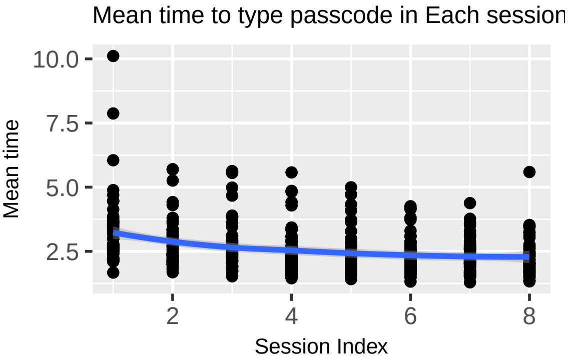
- The dispersion of hold feartures is more than others.
- Up-Down and Down-Down features have minimal standard deviation and high positive kurtosis which can be an indication of outliers.
- The dispersion in hold variables might indicate that the subjects might be varying much in their hold pattern than Up-Down patterns.

Exploratory data analysis : Summary Statistics

```
##
## ========================================================================
## Statistic         N     Mean  St. Dev.  Min   Pctl(25) Pctl(75)  Max
## ------------------------------------------------------------------------
## sessionIndex    20,400 4.500   2.291     1      2.8      6.2      8
## rep             20,400 25.500 14.431     1      13       38       50
## H.period        20,400 0.093   0.030   0.001   0.074    0.108   0.376
## DD.period.t     20,400 0.264   0.221   0.019   0.147    0.306   12.506
## UD.period.t     20,400 0.171   0.227   -0.236  0.050    0.212   12.452
## DD.t.i          20,400 0.169   0.124   0.001   0.114    0.184   4.920
## UD.t.i          20,400 0.083   0.126   -0.162  0.027    0.096   4.800
## H.i             20,400 0.082   0.027   0.003   0.062    0.097   0.331
## DD.i.e          20,400 0.159   0.227   0.001   0.089    0.173   25.987
## UD.i.e          20,400 0.078   0.229   -0.160  0.007    0.093   25.916
## H.e             20,400 0.089   0.031   0.002   0.069    0.103   0.325
## DD.e.five       20,400 0.377   0.265   0.001   0.217    0.457   4.962
## H.five          20,400 0.077   0.022   0.001   0.061    0.091   0.199
## DD.five.Shift.r 20,400 0.439   0.260   0.169   0.308    0.486   8.370
## UD.five.Shift.r 20,400 0.362   0.261   0.086   0.230    0.409   8.291
## H.Shift.r       20,400 0.096   0.034   0.001   0.070    0.117   0.282
## DD.Shift.r.o    20,400 0.251   0.175   0.049   0.156    0.283   4.152
## UD.Shift.r.o    20,400 0.155   0.182   -0.086  0.055    0.191   4.012
## DD.o.a          20,400 0.157   0.107   0.001   0.106    0.168   2.857
## UD.o.a          20,400 0.069   0.109   -0.229  0.017    0.080   2.815
## H.a             20,400 0.106   0.039   0.004   0.082    0.122   2.035
## DD.a.n          20,400 0.151   0.107   0.001   0.096    0.175   3.328
## UD.a.n          20,400 0.044   0.105   -0.236  -0.009   0.069   2.524
## DD.n.l          20,400 0.203   0.150   0.001   0.128    0.229   4.025
## UD.n.l          20,400 0.113   0.160   -0.176  0.024    0.146   3.978
## H.l             20,400 0.096   0.028   0.004   0.077    0.111   0.341
## DD.l.Return     20,400 0.322   0.225   0.008   0.210    0.350   5.884
## UD.l.Return     20,400 0.226   0.231   -0.124  0.114    0.255   5.836
## H.Return        20,400 0.088   0.027   0.003   0.070    0.104   0.265
## ------------------------------------------------------------------------
```

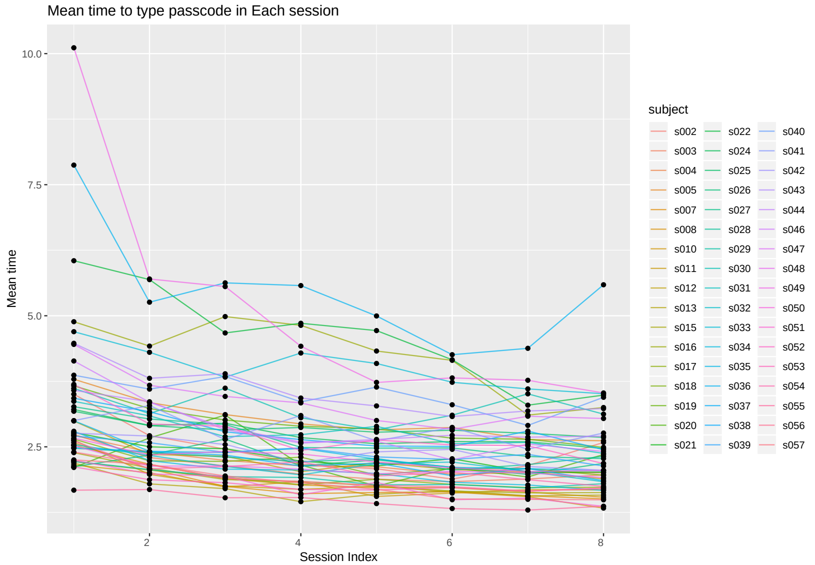
Exploratory data analysis

- The summary statistics show negative minimum values in some UD features. The negative value indicates an overlap i.e the second key is pressed before the first one is released.[2]
- This is an interesting trend to explore.
- First, explore whether the mean time type to passcode changes over sessions.



Exploratory data analysis

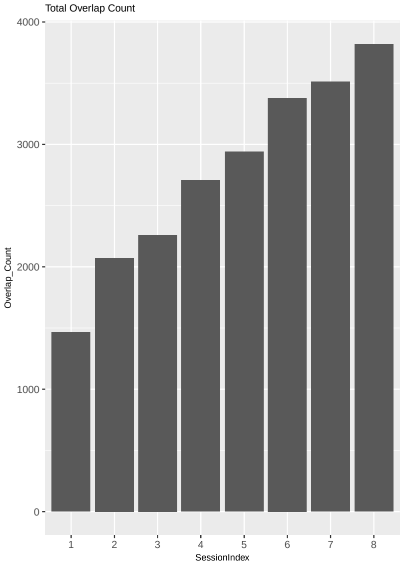
- Subject wise variation is also explored.

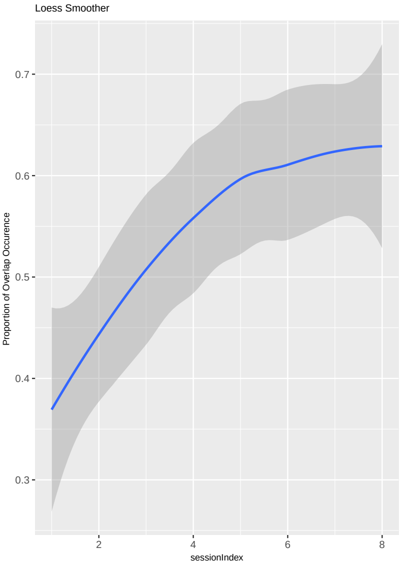


Exploratory data analysis

- On average, the mean time to type passcode seems to be decreasing with each session. However, there are few subjects whose patterns are appearing to be outliers.
- Coming back to the overlaps (Negative values in Up-Down features), exploring the number of overlaps for each session might reveal an interesting trend.

Exploratory data analysis



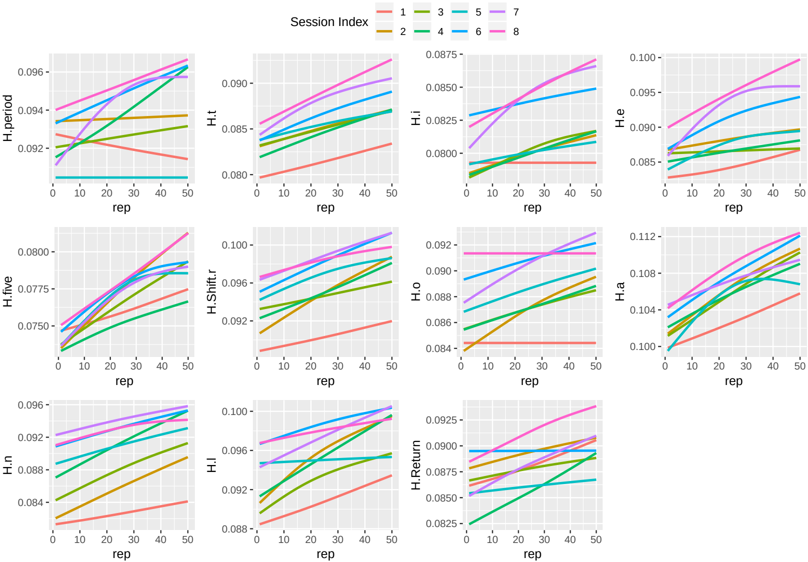


Exploratory data analysis

- The count of overlaps is clearly increasing with each session. The proportion of overlaps is also increasing over sessions.
- The increase in overlaps suggests increase in hold times and decrease in Key-Up-Downs's. (time between holds)
- This can be explored using smoother plots of Hold features and UD features. Since, data is large GAM smoother is used instead of Loess.

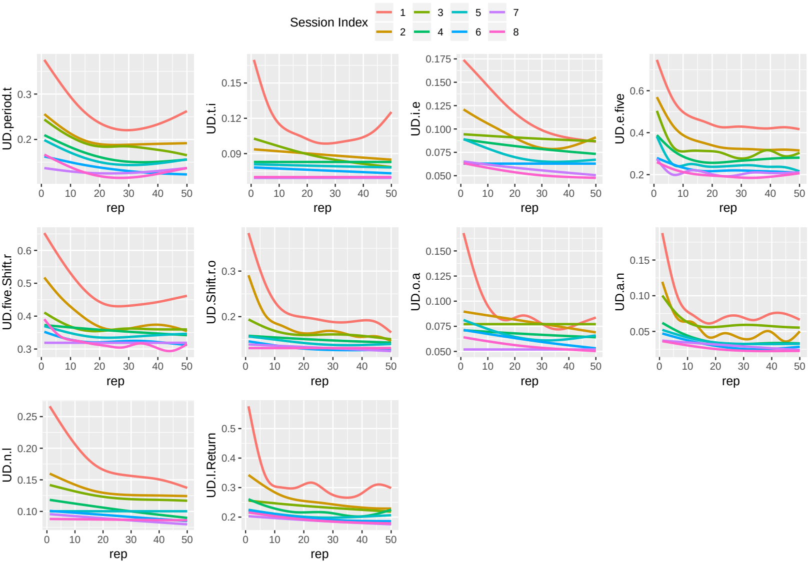
Exploratory data analysis

- Hold times - Increasing over sessions and repetitions.



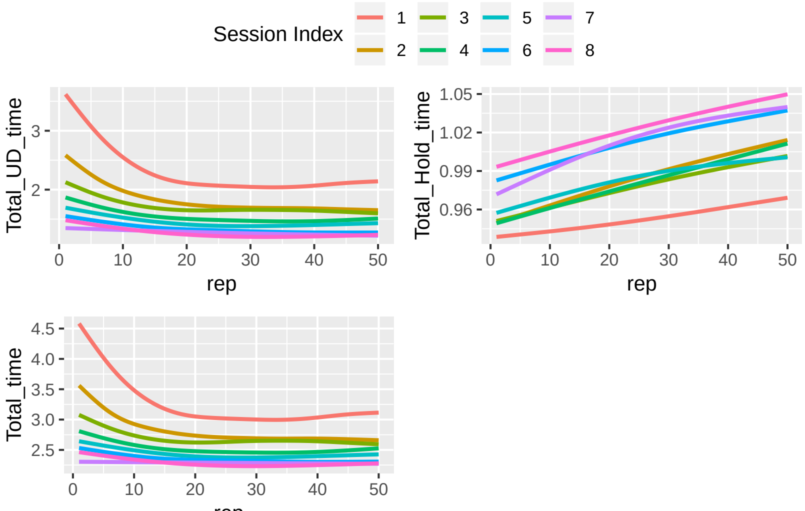
Exploratory data analysis

- UD times - Decreasing over sessions and repetitions.



Exploratory data analysis

- Exploring change in Total Hold time(Sum of Hold features), Total UD time(Sum of UD features) and Total time to type passcode (Sum of DD features and H.Return).



Response Variable Creation

- The total time to type passcode is decreasing with each sessions.
- It is clearly evident that the Hold times increase across sessions and Key-Up-Down times decrease 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. Number of Occurences of Overlap (Overlap_Yes) and No Occurence of overlap (Overlap_No) for each subject in a session are calculated.

The first few rows of the dataset for model looks as follows

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

Response Variable Creation

- Overlap_Yes: Number of occurences of Overlaps for a subject in a session
- Overlap_No: Number of non-occurences of overlaps for a subject in a session
- Both of these will be combined using a two-column matrix and modelled as the response variable[3].
- Since the question of interest is how the typing dynamics change over time, change in patterns of occurences of overlap over sessions can be considered as an appropriate choice of response.
- If the inference through the model that occurence of overlaps increase over sessions is statistically significant, it can be concluded that the conjecture that a person's typing dynamics change over time.

Model Building

- A logistic regression model can be considered with the two individual count observations grouped as the reponse.
- However, a Generalized Linear Model (Logistic Regression) makes assumption of independence of observations. The data used for the model building is a longitudinal data which have repeated measures from the same subjects thus violating the assumption of independence.
- A Logistic **Generalized linear Mixed Effects Model** can be used which adds the random effects to account for the non-independence of observations.
- The primary concern is to explore how keystroke dynamics change over time. 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.

Model Fitting

Generalize Linear Mixed Effects Model with Random Intercept

- The **glmer()** function in the lme4 package is used to fit the Generalized linear Mixed Effects Model.
- sessionIndex independent (explanatory) variable
- grouped observations **Overlap_Yes** and **Overlap_NO** as response
- Subject as random intercept

```
aa <- glmer( cbind(Overlap_Yes,Overlap_No)~sessionIndex
      +(1|subject),family = binomial(),data = anadt)
```

- 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.[4]

Model Fitting

Generalize Linear Mixed Effects Model with Random Slope and Intercept

- sessionIndex independent (explanatory) variable
- grouped observations **Overlap_Yes** and **Overlap_NO** as response
- sessionIndex as random slope and subject as random intercept

```
aa1 <- glmer(cbind(Overlap_Yes,Overlap_No)~sessionIndex +
  (sessionIndex|subject),family = binomial(),data = anadt)
```

Model Fitting

- Two mixed models can be compared using likelihood ratio test.[4]
- The comparison is done to check whether a random slope and intercept model is a significant improvement from random intercept model.
- The p-value indicates that the random intercept and slope model is a significant improvement from the random intercept model.

```
## Data: anadt
## Models:
## aa: cbind(Overlap_Yes, Overlap_No) ~ sessionIndex + (1 | subject)
## aa1: cbind(Overlap_Yes, Overlap_No) ~ sessionIndex + (sessionIndex |
## aa1:     subject)
##     Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
## aa   3 3083.1 3095.1 -1538.5   3077.1
## aa1  5 2666.3 2686.4 -1328.2   2656.3 420.72      2  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Fitted Model : Generalized Linear Mixed Effects Model

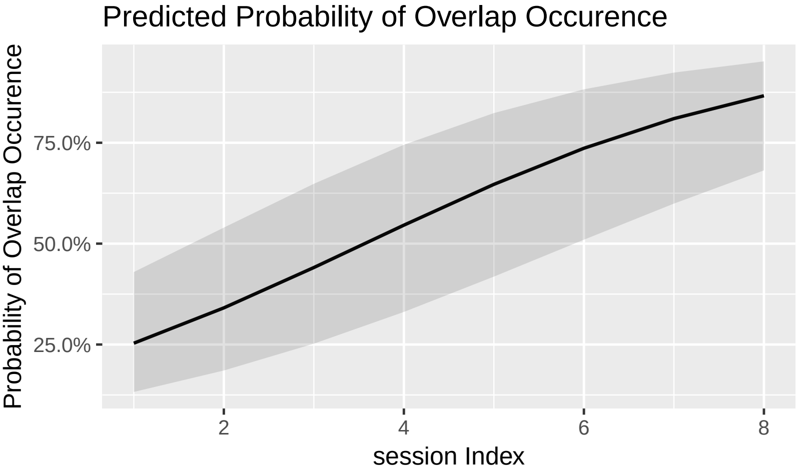
```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula: cbind(Overlap_Yes, Overlap_No) ~ sessionIndex + (sessionIndex |
##     subject)
##    Data: anadt
##
##      AIC      BIC   logLik deviance df.resid
##   2666.3   2686.4  -1328.2   2656.3      403
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -5.3784 -0.9876 -0.0447  0.8758  5.9039
##
## Random effects:
##  Groups  Name         Variance Std.Dev. Corr
##  subject (Intercept)  7.86639  2.8047
##          sessionIndex 0.09617  0.3101   0.13
## Number of obs: 408, groups:  subject, 51
##
## Fixed effects:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.50283    0.40169  -3.741 0.000183 ***
## sessionIndex  0.42163    0.04707   8.958  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr)
## sessionIndx 0.057
```

Inference from fitted Model

- The coefficients of the fixed effects are significant.
- The positive coefficients of sessionIndex suggests that on average for a subject the probability of occurrences of overlap increases over sessions and repetitions.
- 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

Inference from fitted Model

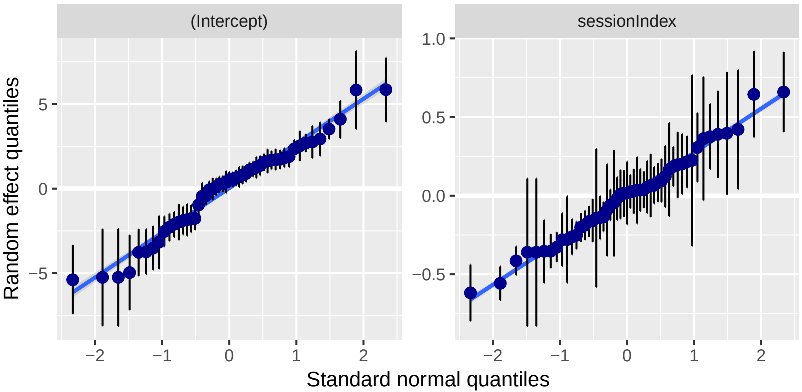
- The predicted probability plot of the fitted Generalized mixed effects model show that the probability of the occurrences of overlaps increases over sessions.



Model Assumptions

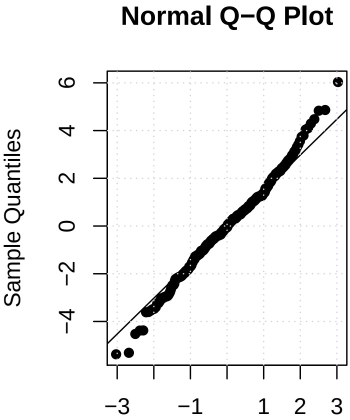
- Q-Q plot of random effects are approximately a straight line. The normality assumption of random effects seems to be satisfied.

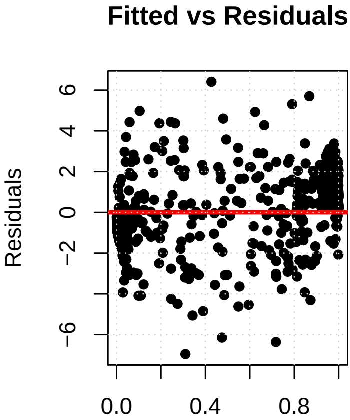
```
## $subject
```



Model Assumptions

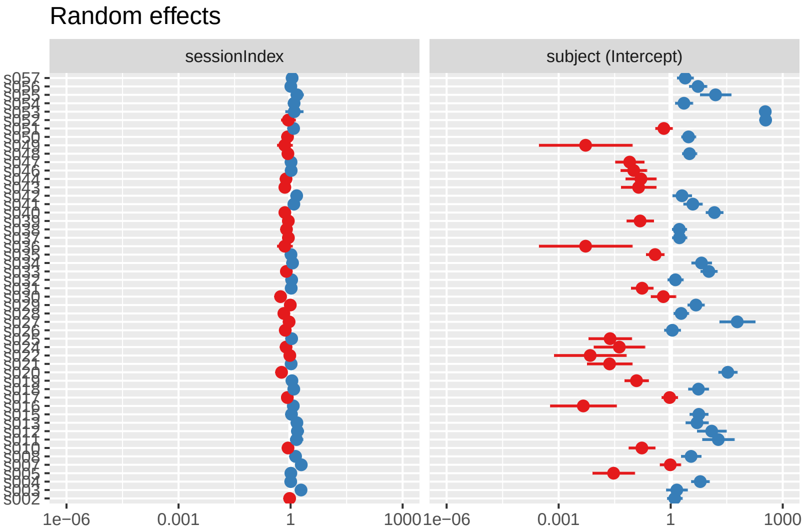
- The Q-Q plot of the residuals approximately looks like a straight line but with slight heavy tails.
- The Residuals vs Fitted plot shows no significant pattern. Homoscedasticity of residuals is satisfied.





Model Assumptions

- The variance due to random intercept (subject) is high compared to random slope (sessionIndex)



Post-Hoc Analysis

Testing for significance of predictors

- A likelihood ratio test is done to test compare significance of predictors.
- The Null model is compared with the full model.
- The random intercept and slope model fitted is a significant improvement from the null model and it has the lowest AIC.

```
## Data: anadt
## Models:
## aa3: cbind(Overlap_Yes, Overlap_No) ~ (sessionIndex | subject)
## aa1: cbind(Overlap_Yes, Overlap_No) ~ sessionIndex + (sessionIndex |
## aa1:     subject)
##     Df    AIC    BIC  logLik deviance Chisq Chi Df Pr(>Chisq)
## aa3  4 2708.8 2724.8 -1350.4   2700.8
## aa1  5 2666.3 2686.4 -1328.2   2656.3 44.47      1  2.583e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Post-Hoc Analysis

Testing for significance of Random Effects

- A glm model without random effects is compared with the glmer model to check significance of random effects.
- The p-value shows that the model with random effects is a significant improvement.

```
## Data: anadt
## Models:
## aa4: cbind(Overlap_Yes, Overlap_No) ~ sessionIndex
## aa1: cbind(Overlap_Yes, Overlap_No) ~ sessionIndex + (sessionIndex |
## aa1:     subject)
##     Df     AIC     BIC  logLik deviance Chisq Chi Df Pr(>Chisq)
## aa4  2 16489.1 16497.2 -8242.6  16485.1
## aa1  5  2666.3  2686.4 -1328.2   2656.3 13829      3  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Conclusions

- The inference that can be drawn from the graphical and statistical analysis performed is that keystroke dynamics for a subject change over time.
- The conclusion from the statistical model built, is that the probability of occurences of overlaps increases over time. Increase in overlaps means that the Hold times on keys increase and latency between holds decreases.
- This indicates that the subjects are able to expect the next key through practice.
- However, the inference is made to generalize the changing trend in keystrokes pattern and the task of detecting outliers is not included in the analysis.

Recommendations

- The Generalized Mixed Effects model built did not model any underlying subset of groups in the sample if they exist.(Ex: Male, Female, Right Handed, Left Handed)
- To identify any sub-groups, clustering can be done.
- If any sub-groups are identified, those sub-groups can be modelled and seen whether the keystoke dynamics significantly differ between those groups.

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] HOTHORN, T. (2017). HANDBOOK OF STATISTICAL ANALYSES USING R, THIRD EDITION. S.l.: CRC PRESS.

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

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

[6] Hlavac, Marek (2018). stargazer: Well-Formatted Regression and Summary Statistics Tables. R package version 5.2.2. https://CRAN.R-project.org/package=stargazer

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

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

[9] http://ddar.datavis.ca/pages/extra/titanic-glm-ex.pdf

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

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