Estimating Whether People's Typing Dynamics Change Over Time

**Charts mentioned can be found in the PowerPoint slides with the corresponding heading and can be reproduced via the annotated code with the corresponding heading*

Contents

Objective

For this project, I'm utilizing a dataset used in a previous study by Carnegie Melon University where different machine learning models were tested to try to find a way to discriminate between users via their typing rhythms. The main goal of that study was to find a machine learning model with low enough error rates to obtain European Union approval so that it could be used legally and with partnership from European governments. The fields in the dataset include 51 subjects, 8 sessions per subject, 50 repetitions of typing the password: " \(.tie5Roanl\)", and 31 fields for actions required to type the password. Also, this dataset simulates typing the same password multiple times(50 in this case) in a given day, represented by repetition number, and 8 days in a row, represented by session number. (Killourhy & Maxion, 2009)

I'm starting with a null hypothesis that a person's typing dynamics change over time, short term and long term, which I'm going to attempt to reject. To do this, I will analyze the dataset via a mixed model approach. The response variable I am interested in here is total time to enter the password, while I am interested in the effects of the variables, rep which is the times the password was repeatedly typed in a day, session which is the days of the continuous repetitions of the password, and subject, which is the person typing the password. For this project, rep and session are fixed effects and subject is the nested random effect I will be testing, since rep and session are consistent actions but subject varies depending on who was available to be tested at the time.

Assumptions with this dataset are that it has been cleaned and verified of any typos and entry errors.

Exploratory Analysis

My exploratory analysis includes box plots showing the average improvement of the subjects between their first and last sessions as well as first and last repetition per session. These plots show that there are outliers in the response variable of the dataset where some subjects committed actions during the typing of the password that were much slower or faster than they usually would, or in some cases, much slower or faster than other subjects. Furthermore, these plots show that on average, there is a change in speed, almost always an improvement, between consecutive sessions and between consecutive repetitions. In fact, between the first and last sessions of the subjects, there is an average improvement of almost 2 seconds. Finally, I created a histogram overlayed with a density plot to check whether the data is normally distributed, which it seems to not be as it's too heavily right-skewed.

For this project, as the outliers can be attributed to the randomness of the subjects, since people often make mistakes when typing ambiguous passwords, and the data is assumed to have been cleaned and verified, I have decided not to remove outliers.

Furthermore, I decided to use penalized quasi-likelihood along with random intercept and random intercept and slope models to model the data. 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. (Everitt & Hothorn, 2014) More common methods such as restricted maximum likelihood and maximum likelihood require that the data be normally distributed. (Pilowsky, 2018) Penalized quasi-likelihood just requires that the response variable not have a mean less than five and the response variable does not fit a discrete count distribution such as Poisson or Binomial, or that the response variable is not binary. (Pilowsky, 2018) The response variable's mean is greater than 5 and is not binary so this technique should work.

Penalized Quasi-Likelihood:

$$\hat{g}_n \in \arg\max_{g \in \delta} [\hat{Q}_n(F(g)) - \lambda_n^2 J^2(g)]$$

(Mammen & van de Gee, 1997)

Random intercept and random intercept and slope models are commonly used as opposed to generalized linear models for datasets with repeated measurements. (Everitt & Hothorn, 2014) These models assume that correlation amongst independent repeated measurements on the same unit arises from the shared unobserved variables and that time has a fixed effect. (Everitt & Hothorn, 2014) The difference between random intercept and random intercept and slope models are that random intercept models measure dissimilarity in intercepts while random intercept and slope models measure dissimilarity in intercepts and slopes. (Everitt & Hothorn, 2014)

Random Intercept model:

$$y_{ij} = \beta_0 + \beta_1 t_j + u_i + \varepsilon_{ij}$$

(Everitt & Hothorn, 2014)

Random Intercept and Slope model:

$$y_{ij} = \beta_0 + \beta_1 t_j + u_i + v_i t_j + \varepsilon_{ij}$$ (Everitt & Hothorn, 2014)

Next I utilized quantile comparison plots to check what probability distribution best fits the response variable. I tested with Normal, Lognormal, Gamma, Negative Binomial and Poisson. I decided to fit the dataset to a Gamma distribution as although Poisson seemed like a slightly closer fit, it also would require a transformation of the response variable from continuous to discrete, or in other words rounding to whole numbers.

Statistical Analysis

Here I fitted the dataset to a gamma model via penalized quasi-likelihood. I tested fit with a random intercept model as well as a random intercept and slope model. (Arbor Custom Analytics, 2020) Also, these were tested with identity and inverse link functions. The GLMMPQL function does not output AIC, BIC, or Log Likelihood and thus I tested fit via residuals charts. These charts showed that the random intercept model with the identity link function and the random intercept model with the identity link function are tied or close to tied for having the closest fit to the dataset.

The results of these two models show that repetitions and consecutive sessions do influence password typing speed. In fact, the p value obtained shows that the null hypothesis has not been disproven. Results show that total time taken to type the password decreases as repetitions increase and to a greater extent as session increases. In other words, people tend to type a password faster the more they type it in a given day and when repeatedly typing the password on consecutive days.

Post Hoc Analysis

For the post hoc analysis, I decided to further confirm whether the response variable is not normally distributed and compare the fixed effects to determine if there's an interaction between them. I tested this via a Shapiro-Wilkins test for normality grouped by subject as this test has a limit of 3500 for observations, and by manipulating the closest fitting model to have a fixed variable of session:repetition. The results of this analysis were that the data is definitely not of a normal distribution and that there is an interaction between session and repetition although it has a very small effect on the response variable compared to that of session and repetition individually.

Conclusions

The random effects model I created that fits the data the best indicates that the time taken to type a password, changes over time as one retypes it throughout the day and on consecutive days. Thus, I've failed to reject the null hypothesis and I have confirmed that at least within the boundaries of this sample and to a statistically significant extent, a person's typing dynamics change over time, short and long term.

Also, the results of the post-hoc analysis showed that the data is extremely likely to not be normally distributed and that there is an interaction between session and repetition.

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.

Killourhy, K. S., & Maxion, R. A. (2009). *Comparing Anomaly-Detection Algorithms for Keystroke Dynamics*. Retrieved from Carnegie Melon University School of Computer Science: https://www.cs.cmu.edu/~keystroke/KillourhyMaxion09.pdf

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

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