Evaluating Free Rides in Set Visualizations: Overall Analysis

Tuesday 05 March, 2019

In this document the groups are referred to in the following way:

1. T only: Text
2. L&T: Linear
3. V&T: Venn, and
4. E&T: Euler.

Methods and Software

Statistical analysis was based on the Generalized Estimating Equations method (Liang and Zeger 1986) as this is implemented in the R (R Core Team 2018) package geepack (Højsgaard, Halekoh, and Yan 2005). In addition, the function `ComparisonStats` was developed to evaluate the statistical significance of the desired comparisons for the accuracy and time data.

```
> ComparisonStats <- function(FittedModel, Lmatrix, alpha = 0.05) {
+     Lmatrix <- matrix(Lmatrix, nrow = 1)
+     ModelBetas <- FittedModel$geese$beta
+     ModelVCov <- FittedModel$geese$vbeta
+     Estimate <- drop(Lmatrix %*% ModelBetas)
+     SdError <- sqrt(drop(Lmatrix %*% ModelVCov %*% t(Lmatrix)))
+     CBs <- Estimate + qnorm(c(alpha/2, 1 - alpha/2)) * SdError
+     pvalue <- 2 * pnorm(-abs(Estimate/SdError))
+     ans <- c(exp(c(Estimate, CBs)), round(pvalue, 4))
+     names(ans) <- c("Estimate", paste0((1 - alpha) * 100, "% LB"), paste0((1 -
+         alpha) * 100, "% UB"), "p-value")
+     ans
+ }
```

Import Data

The full data were imported by executing the following commands:

```
> free_rides <- read.csv("./data/main.csv")
> names(free_rides) <- gsub("[.]", "_", make.names(names(free_rides), unique = TRUE))
```

Analysis of accuracy data

The following regression model was fitted to the accuracy data

$$\log\left[\frac{\Pr(Y_{ij}=1)}{1-\Pr(Y_{ij}=1)}\right] = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}$$

where

- $\Pr(Y_{ij} = 1)$ is the probability for participant i to answer question j correctly.
- x_{i1} is the indicator for the *Linear* treatment,
- x_{i2} is the indicator for the *Text* Treatment,
- x_{i3} is the indicator for the *Venn* Treatment,

for $i = 1, \ldots, 404$, corresponding to the individual participants, and $j = 1, \ldots, 20$ corresponding to the questions (questions $6 - 21$ are labeled as $5, \ldots, 20$) respectively.

```
> library(geepack)
> reduced_model <- geeglm(formula = Correct ~ Treatment, id = Study_Id, data = free_rides,
+     family = binomial)
```

Comparison of treatments

Euler = *Linear* > *Venn* = *Text*.

```
> ## Linear versus Text
> ComparisonStats(reduced_model, c(0, 1, -1, 0))
 Estimate    95% LB    95% UB   p-value
 6.742312  4.525890 10.044161  0.000000
> ## Linear versus Venn
> ComparisonStats(reduced_model, c(0, 1, 0, -1))
 Estimate    95% LB    95% UB   p-value
 6.753965  4.368747 10.441447  0.000000
> ## Linear versus Euler
> ComparisonStats(reduced_model, c(0, 1, 0, 0))
 Estimate    95% LB    95% UB   p-value
0.6954797 0.4222735 1.1454470 0.1537000
> ## Text versus Venn
> ComparisonStats(reduced_model, c(0, 0, 1, -1))
 Estimate    95% LB    95% UB   p-value
1.0017283 0.7518112 1.3347229 0.9906000
> ## Text versus Euler
> ComparisonStats(reduced_model, c(0, 0, 1, 0))
  Estimate     95% LB     95% UB    p-value
0.10315152 0.07081024 0.15026406 0.00000000
> ## Venn versus Euler
> ComparisonStats(reduced_model, c(0, 0, 0, 1))
  Estimate     95% LB     95% UB    p-value
0.10297354 0.06797965 0.15598124 0.00000000
```

Analysis of time data

The following regression model was fitted to the time data

$$\log(Z_{ij}) = \gamma_0 + \gamma_1 x_{i1} + \gamma_2 x_{i2} + \gamma_3 x_{i3}$$

where

- Z_{ij} is the time needed for participant i to answer question j correctly.
- x_{i1} is the indicator for the *Linear* treatment,
- x_{i2} is the indicator for the *Text* Treatment,
- x_{i3} is the indicator for the *Venn* Treatment,

for $i = 1, \ldots, 404$, corresponding to the individual participants, and $j = 1, \ldots, 20$ corresponding to the questions (questions $6 - 21$ are labeled as $5, \ldots, 20$) respectively.

```
> library(geepack)
> free_rides$log_time <- log(free_rides$Time)
> reduced_model <- geeglm(formula = log_time ~ Treatment,
+                         id = Study_Id, data = free_rides[free_rides$Correct == 1,])
```

Comparison of treatments

Euler = Linear > Venn = Text.

```
> ## Linear versus Text
> ComparisonStats(reduced_model, c(0, 1, -1, 0))
 Estimate    95% LB    95% UB   p-value
0.6182399 0.5510422 0.6936321 0.0000000
> ## Linear versus Venn
> ComparisonStats(reduced_model, c(0, 1, 0, -1))
 Estimate    95% LB    95% UB   p-value
0.6265788 0.5587246 0.7026737 0.0000000
> ## Linear versus Euler
> ComparisonStats(reduced_model, c(0, 1, 0, 0))
 Estimate    95% LB    95% UB   p-value
0.9499680 0.8531786 1.0577377 0.3492000
> ## Text versus Venn
> ComparisonStats(reduced_model, c(0, 0, 1, -1))
Estimate   95% LB   95% UB  p-value
1.013488 0.891809 1.151769 0.837300
> ## Text versus Euler
> ComparisonStats(reduced_model, c(0, 0, 1, 0))
Estimate   95% LB   95% UB  p-value
1.536569 1.360734 1.735125 0.000000
> ## Venn versus Euler
> ComparisonStats(reduced_model, c(0, 0, 0, 1))
Estimate   95% LB   95% UB  p-value
1.516119 1.343192 1.711308 0.000000
```

References

Højsgaard, S., U. Halekoh, and J. Yan. 2005. "The R Package geepack for Generalized Estimating Equations." *Journal of Statistical Software* 15: 1–11.

Liang, K.Y., and S.L. Zeger. 1986. "Longitudinal Data Analysis Using Generalized Linear Models." *Biometrika* 73 (1): 13–22.

R Core Team. 2018. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org/.