Observational Advantages – Overall Analysis – No question types

Monday 21 October, 2019

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 2019) 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:

```
> library(readxl)
> free_rides <- read_excel("data/AllMainStudy_Data.xlsx", col_types = c("numeric",
+     "skip", "text", "text", "numeric", "text", "numeric", "numeric", "numeric",
+     "numeric", "text", "text", "text", "numeric", "text", "text", "numeric"))
> names(free_rides) <- gsub("[.]", "_", make.names(names(free_rides), unique = TRUE))
> free_rides <- free_rides[free_rides$Study_Id != 392, ]
```

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 $i = 1, \ldots, 418$, 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
 7.904064  5.406137 11.556167  0.000000
> ## Linear versus Venn
> ComparisonStats(reduced_model, c(0, 1, 0, -1))
Estimate   95% LB   95% UB  p-value
5.662058 3.756341 8.534610 0.000000
> ## Linear versus Euler
> ComparisonStats(reduced_model, c(0, 1, 0, 0))
 Estimate    95% LB    95% UB   p-value
0.8701159 0.5404592 1.4008489 0.5669000
> ## Text versus Venn
> ComparisonStats(reduced_model, c(0, 0, 1, -1))
 Estimate    95% LB    95% UB   p-value
0.7163478 0.5395637 0.9510538 0.0211000
> ## Text versus Euler
> ComparisonStats(reduced_model, c(0, 0, 1, 0))
  Estimate     95% LB     95% UB    p-value
0.11008462 0.07585397 0.15976255 0.00000000
> ## Venn versus Euler
> ComparisonStats(reduced_model, c(0, 0, 0, 1))
 Estimate    95% LB    95% UB   p-value
0.1536748 0.1026515 0.2300595 0.0000000
```

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 $i = 1, \ldots, 418$, 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
0.5287369 0.4731198 0.5908920 0.0000000
> ## Linear versus Venn
> ComparisonStats(reduced_model, c(0, 1, 0, -1))
 Estimate    95% LB    95% UB   p-value
0.6295972 0.5749555 0.6894319 0.0000000
> ## Linear versus Euler
> ComparisonStats(reduced_model, c(0, 1, 0, 0))
 Estimate    95% LB    95% UB   p-value
0.9313072 0.8474650 1.0234443 0.1393000
> ## Text versus Venn
> ComparisonStats(reduced_model, c(0, 0, 1, -1))
Estimate   95% LB   95% UB  p-value
1.190757 1.066625 1.329335 0.001900
> ## Text versus Euler
> ComparisonStats(reduced_model, c(0, 0, 1, 0))
Estimate   95% LB   95% UB  p-value
1.761381 1.573121 1.972172 0.000000
> ## Venn versus Euler
> ComparisonStats(reduced_model, c(0, 0, 0, 1))
Estimate   95% LB   95% UB  p-value
1.479211 1.347716 1.623536 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. 2019. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org/.