Free Rides – Analysis By Task Type

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 functions `ComparisonStats` and `AllContrasts` were 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
+ }
```

```
> contrasts_of_interest <- function(model, data, ignore_levels = NULL, varying = NULL) {
+     data <- as.data.frame(data)
+     xformula <- formula(model)
+     model_terms <- terms(xformula)
+     model_factors <- rownames(attr(model_terms, "factors"))[-1]
+     data_contrasts <- expand.grid(lapply(data[, model_factors], unique))
+     xformula <- as.formula(paste("~", paste(attr(terms(xformula), "term.labels"),
+         collapse = "+")))
+     model_frame_contrasts <- model.matrix(xformula, data_contrasts)
+     if (!is.null(ignore_levels)) {
+         stopifnot(all(names(ignore_levels) %in% names(data_contrasts)))
+         ignore_levels_index <- which(rowSums(data_contrasts == ignore_levels) !=
+             0)
+         data_contrasts <- data_contrasts[-ignore_levels_index, ]
+         model_frame_contrasts <- model_frame_contrasts[-ignore_levels_index,
+             ]
+     }
+     names_contrasts <- do.call("paste", c(data_contrasts, sep = " & "))
+     if (!is.null(varying)) {
```

```
+         vunique <- unique(data_contrasts[, varying])
+         p <- nrow(model_frame_contrasts)/length(vunique)
+     }
+     if ((is.null(varying)) | (p == 1)) {
+         for (i in 1:(nrow(data_contrasts) - 1)) {
+             contrast_1 <- model_frame_contrasts[i, ]
+             for (j in (i + 1):nrow(data_contrasts)) {
+                 contrast_2 <- model_frame_contrasts[j, ]
+                 cat("## (", names_contrasts[i], ") versus (", names_contrasts[j],
+                   ")\n", sep = "")
+                 print(ComparisonStats(model, contrast_1 - contrast_2))
+             }
+         }
+     } else {
+         for (i in 1:p) {
+             contrast_1 <- model_frame_contrasts[i, ]
+             for (j in 2:length(vunique)) {
+                 contrast_2 <- model_frame_contrasts[p * (j - 1) + i, ]
+                 cat("## (", names_contrasts[i], ") versus (", names_contrasts[p *
+                   (j - 1) + i], ")\n", sep = "")
+                 print(ComparisonStats(model, contrast_1 - contrast_2))
+             }
+         }
+     }
+ }
```

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

$$\begin{aligned}\log\left[\frac{\Pr(Y_{ij}=1)}{1-\Pr(Y_{ij}=1)}\right] =& \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} \\ &+ \beta_5 x_{i1} x_{i4} + \beta_6 x_{i1} x_{i4} + \beta_7 x_{i3} x_{i4}\end{aligned}$$

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,
- x_{i4} is the indicator for the *subset* Question Type,

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

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

Comparing this model to the model with no interaction term

```
> reduced_model <- geeglm(formula = Correct ~ factor(Treatment) + factor(Question_Type),
+     id = Study_Id, data = free_rides, family = binomial)
> Waldts <- anova(full_model, reduced_model)
> Waldts
Analysis of 'Wald statistic' Table

Model 1 Correct ~ Treatment * Question_Type
Model 2 Correct ~ factor(Treatment) + factor(Question_Type)
  Df     X2 P(>|Chi|)
1  3 29.115 2.118e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

we can conclude that at least one of the interplay of *Treatment* and *Question Type* is significant (p - value $= 2.1182825 \times 10^{-6}$).

Comparison of treatments for the same question type

For *Question_Type* = *disjoint*: *Euler* = *Linear* > *Venn* > *Text*.

For *Question_Type* = *subset*: *Euler* = *Linear* > *Text* > *Venn*.

Question Type: disjoint

```
> contrasts_of_interest(full_model, free_rides, ignore_levels = list(Question_Type = "subset"),
+     varying = "Treatment")
## (Linear & disjoint) versus (Text & disjoint)
 Estimate    95% LB    95% UB   p-value
 9.627826  6.121943 15.141439  0.000000
## (Linear & disjoint) versus (Venn & disjoint)
 Estimate    95% LB    95% UB   p-value
 6.427566  3.934650 10.499941  0.000000
## (Linear & disjoint) versus (Euler & disjoint)
 Estimate    95% LB    95% UB   p-value
0.7440860 0.4221224 1.3116196 0.3068000
## (Text & disjoint) versus (Venn & disjoint)
 Estimate    95% LB    95% UB   p-value
0.6676030 0.4837641 0.9213040 0.0139000
## (Text & disjoint) versus (Euler & disjoint)
  Estimate     95% LB     95% UB    p-value
0.07728495 0.05031443 0.11871272 0.00000000
## (Venn & disjoint) versus (Euler & disjoint)
  Estimate     95% LB     95% UB    p-value
0.11576483 0.07241725 0.18505943 0.00000000
```

Question Type: subset

```
> contrasts_of_interest(full_model, free_rides, ignore_levels = list(Question_Type = "disjoint"),
+     varying = "Treatment")
## (Linear & subset) versus (Text & subset)
Estimate   95% LB   95% UB  p-value
4.707814 3.025777 7.324898 0.000000
## (Linear & subset) versus (Venn & subset)
 Estimate    95% LB    95% UB   p-value
 7.098631  4.424034 11.390185  0.000000
## (Linear & subset) versus (Euler & subset)
 Estimate    95% LB    95% UB   p-value
0.6459836 0.3753130 1.1118584 0.1147000
## (Text & subset) versus (Venn & subset)
Estimate   95% LB   95% UB  p-value
1.507840 1.072838 2.119222 0.018000
## (Text & subset) versus (Euler & subset)
  Estimate     95% LB     95% UB    p-value
0.13721520 0.08902935 0.21148096 0.00000000
## (Venn & subset) versus (Euler & subset)
  Estimate     95% LB     95% UB    p-value
0.09100116 0.05721776 0.14473147 0.00000000
```

Comparison of question types within treatment

For *Treatment* = *Euler*, *Linear* and *Venn*: *subset* = *disjoint*.

For *Treatment* = *Text*: *subset* > *disjoint*.

```
> contrasts_of_interest(full_model, free_rides, varying = "Question_Type")
## (Linear & subset) versus (Linear & disjoint)
 Estimate    95% LB    95% UB   p-value
1.0378350 0.7413309 1.4529295 0.8287000
## (Text & subset) versus (Text & disjoint)
Estimate   95% LB   95% UB  p-value
2.122449 1.716693 2.624109 0.000000
## (Venn & subset) versus (Venn & disjoint)
 Estimate    95% LB    95% UB   p-value
0.9397238 0.7402322 1.1929781 0.6096000
## (Euler & subset) versus (Euler & disjoint)
 Estimate    95% LB    95% UB   p-value
1.1954459 0.8394139 1.7024866 0.3224000
```

Analysis of time data

The following regression model was fitted to the time data

$$\begin{aligned}\log(Z_{ij}) =& \gamma_0 + \gamma_1 x_{i1} + \gamma_2 x_{i2} + \gamma_3 x_{i3} + \gamma_4 x_{i4} \\ &+ \gamma_5 x_{i1} x_{i4} + \gamma_6 x_{i1} x_{i4} + \gamma_7 x_{i3} x_{i4}\end{aligned}$$

where

- Z_{ij} is the time participant i needed 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,
- x_{i4} is the indicator for the *subset* Question Type,

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

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

Comparing this model to the model with no interaction term

```
> reduced_model <- geeglm(formula = log_time ~ Treatment + Question_Type, id = Study_Id,
+     data = free_rides[free_rides$Correct == 1, ])
> Waldts <- anova(reduced_model, full_model)
> Waldts
Analysis of 'Wald statistic' Table

Model 1 log_time ~ Treatment * Question_Type
Model 2 log_time ~ Treatment + Question_Type
  Df     X2 P(>|Chi|)
1  3 14.703  0.002089 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

we can conclude that at least one of the interplay of *Treatment* and *Question Type* is significant (p - value $= 0.0020892$).

Comparison of treatments for the same question type

For *Question_Type* = *disjoint*: *Euler* = *Linear* > *Venn* = *Text*.

For *Question_Type* = *subset*: *Euler* = *Linear* > *Venn* = *Text*.

Question Type: disjoint

```
> contrasts_of_interest(full_model, free_rides, ignore_levels = list(Question_Type = "subset"),
+     varying = "Treatment")
## (Linear & disjoint) versus (Text & disjoint)
 Estimate    95% LB    95% UB   p-value
0.5679305 0.4977509 0.6480051 0.0000000
## (Linear & disjoint) versus (Venn & disjoint)
 Estimate    95% LB    95% UB   p-value
0.6380435 0.5646284 0.7210042 0.0000000
## (Linear & disjoint) versus (Euler & disjoint)
 Estimate    95% LB    95% UB   p-value
0.9715226 0.8674176 1.0881219 0.6174000
## (Text & disjoint) versus (Venn & disjoint)
Estimate   95% LB   95% UB  p-value
1.123453 0.971630 1.299000 0.116100
## (Text & disjoint) versus (Euler & disjoint)
```

```
Estimate   95% LB   95% UB  p-value
1.710636 1.490457 1.963341 0.000000
## (Venn & disjoint) versus (Euler & disjoint)
Estimate   95% LB   95% UB  p-value
1.522659 1.338961 1.731560 0.000000
```

Question Type: subset

```
> contrasts_of_interest(full_model, free_rides, ignore_levels = list(Question_Type = "disjoint"),
+     varying = "Treatment")
## (Linear & subset) versus (Text & subset)
 Estimate    95% LB    95% UB   p-value
0.6559027 0.5809397 0.7405388 0.0000000
## (Linear & subset) versus (Venn & subset)
 Estimate    95% LB    95% UB   p-value
0.6149487 0.5451478 0.6936870 0.0000000
## (Linear & subset) versus (Euler & subset)
 Estimate    95% LB    95% UB   p-value
0.9292354 0.8327335 1.0369205 0.1896000
## (Text & subset) versus (Venn & subset)
 Estimate    95% LB    95% UB   p-value
0.9375609 0.8185926 1.0738190 0.3517000
## (Text & subset) versus (Euler & subset)
Estimate   95% LB   95% UB  p-value
1.416727 1.248790 1.607249 0.000000
## (Venn & subset) versus (Euler & subset)
Estimate   95% LB   95% UB  p-value
1.511078 1.333088 1.712832 0.000000
```

Comparison of question types within treatment

For *Linear* and *Venn*: *subset* = *disjoint*.

For *Text* and *Euler*: *disjoint* > *subset*.

```
> contrasts_of_interest(full_model, free_rides, varying = "Question_Type")
## (Linear & subset) versus (Linear & disjoint)
 Estimate    95% LB    95% UB   p-value
1.0024691 0.9580399 1.0489588 0.9151000
## (Text & subset) versus (Text & disjoint)
 Estimate    95% LB    95% UB   p-value
0.8680141 0.7922987 0.9509652 0.0024000
## (Venn & subset) versus (Venn & disjoint)
 Estimate    95% LB    95% UB   p-value
1.0401174 0.9735823 1.1111995 0.2435000
## (Euler & subset) versus (Euler & disjoint)
Estimate   95% LB   95% UB  p-value
1.048089 1.008108 1.089655 0.017900
```

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/.