Report coursework assignment A - 2021

CS4125 Seminar Research Methodology for Data Science

Nikki Bouman (4597648), Anuj Singh (), Gwennan Smitskamp ()

20/04/2021

Contents

1 Part 1 - Design and set-up of true experiment

1.1 The motivation for the planned research

(Max 250 words)

1.2 The theory underlying the research

(Max 250 words) Preferable based on theories reported in literature

1.3 Research questions

The research question that will be examined in the experiment (or alternatively the hypothesis that will be tested in the experiment)

1.4 The related conceptual model

This model should include: *Independent variable(s)* Dependent variable *Mediating variable (at least 1)* Moderating variable (at least 1)

1.5 Experimental Design

Note that the study should have a true experimental design

1.6 Experimental procedure

Describe how the experiment will be executed step by step

1.7 Measures

Describe the measure that will be used

1.8 Participants

Describe which participants will recruit in the study and how they will be recruited

1.9 Suggested statistical analyses

Describe the statistical test you suggest to care out on the collected data

2 Part 2 - Generalized linear models

2.1 Question 1 Twitter sentiment analysis (Between groups - single factor)

2.1.1 Conceptual model

Make a conceptual model for the following research question: Is there a difference in the sentiment of the tweets related to the different celebrities?

2.1.2 Collecting tweets, and data preparation

Include the annotated R script (excluding your personal Keys and Access Tokens information), but put echo=FALSE, so code is not included in the output pdf file.

2.1.3 Homogeneity of variance analysis

Analyze the homogeneity of variance of sentiments of the tweets of the different celebrities, and provide interpretation

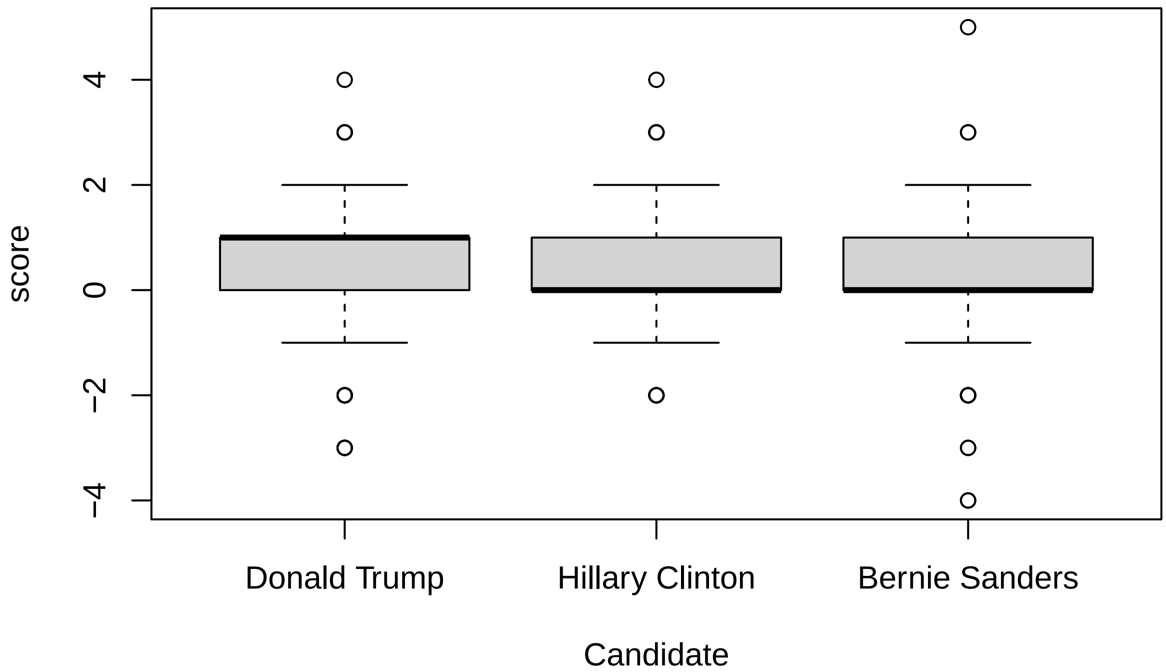
```
#include your code and output in the document
leveneTest(semFrame$score, semFrame$Candidate, center = median)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value   Pr(>F)
## group   2  5.2913 0.005409 **
##       384
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

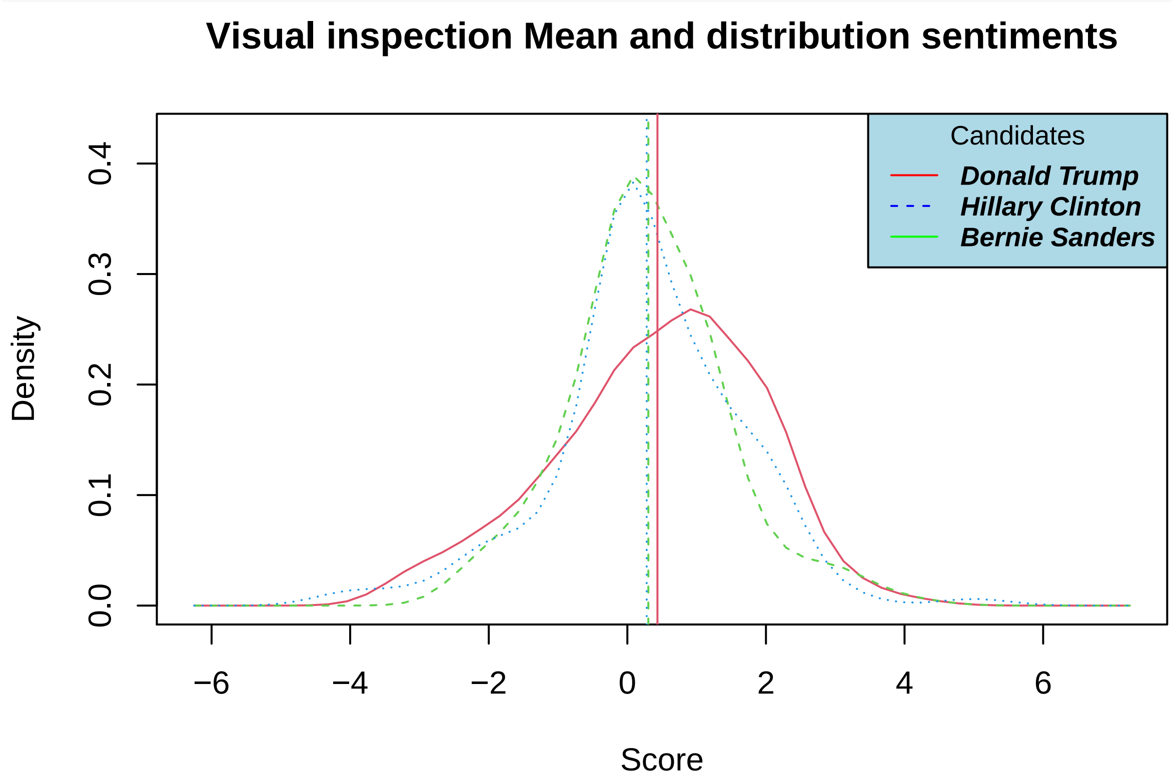
2.1.4 Visual inspection Mean and distribution sentiments

Graphically examine the mean and distribution sentiments of tweets for each celebrity, and provide interpretation

```
boxplot(score ~ Candidate, data = semFrame)
```




```
sm.density.compare(semFrame$score, semFrame$Candidate, xlab = "Score")
  title(main="Visual inspection Mean and distribution sentiments")
  legend('topright', legend = levels(semFrame$Candidate), col=c('red', 'blue', 'green'),
         title="Candidates", lty=1:2, cex=0.8, text.font = 4, bg='lightblue')
means <- aggregate(semFrame$score ~ semFrame$Candidate, FUN = mean)
abline(v = means[1,2], col = 2)
abline(v = means[2,2], col = 3, lty = 2)
abline(v = means[3,2], col = 4, lty = 3)
```



```
#include your code and output in the document
```

2.1.5 Frequentist approach

2.1.5.1 Linear model Use a linear model to analyze whether the knowledge to which celebrity a tweet relates has a significant impact on explaining the sentiments of the tweets. Provide interpretation of results

```
model0 <- lm(formula = score ~ 1 , data = semFrame)
model1 <- lm(formula = score ~ Candidate , data = semFrame)
anova(model0, model1, test = "F")
```

```
## Analysis of Variance Table
##
## Model 1: score ~ 1
## Model 2: score ~ Candidate
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    386 654.66
## 2    384 652.85  2    1.8036 0.5304 0.5888
```

```
summary(model1)
```

```
##
## Call:
## lm(formula = score ~ Candidate, data = semFrame)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.2791 -0.4341 -0.2791  0.6977  4.7209
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)                0.4341     0.1148   3.781 0.000181 ***
## CandidateHillary Clinton  -0.1318     0.1624  -0.812 0.417465
## CandidateBernie Sanders   -0.1550     0.1624  -0.955 0.340207
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.304 on 384 degrees of freedom
## Multiple R-squared:  0.002755,   Adjusted R-squared:  -0.002439
## F-statistic: 0.5304 on 2 and 384 DF,  p-value: 0.5888
```

```
smodel0 <-summary(model0)
llm0 <-sum(dnorm(semFrame$score, mean = predict(model0), sd= smodel0$sigma, log=TRUE))
AIC_model0 <- -2*llm0 + 2*2
AIC_model0
```

```
## [1] 1305.7
```

```
smodel1 <-summary(model1)
llm1 <-sum(dnorm(semFrame$score, mean = predict(model1), sd= smodel1$sigma, log=TRUE))
AIC_model1 <- -2*llm1 + 2*3
AIC_model1
```

```
## [1] 1306.643
```

```
#include your code and output in the document
```

2.1.5.2 Post Hoc analysis

If a model that includes the celebrity is better in explaining the sentiments of tweets than a model without such predictor, conduct a post-hoc analysis with e.g. Bonferroni correction, to examine which of celebrity tweets differ from the other celebrity tweets. Provide interpretation of the results

```
pairwise.t.test(semFrame$score, semFrame$Candidate, paired = FALSE, p.adjust.method = "bonferroni")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  semFrame$score and semFrame$Candidate
##
##                 Donald Trump Hillary Clinton
## Hillary Clinton 1            -
## Bernie Sanders  1            1
##
## P value adjustment method: bonferroni
```

```
#include your code and output in the document
```

2.1.5.3 Report section for a scientific publication

Write a small section for a scientific publication, in which you report the results of the analyses, and explain the conclusions that can be drawn.

2.1.6 Bayesian Approach

2.1.6.1 Model description

Describe the mathematical model fitted on the most extensive model. (hint, look at the mark down file of the lectures to see example on formulate mathematical models in markdown). Justify the priors.

Comparing multiple levels

2.1.6.2 Model comparison

Conduct model analysis and provide brief interpretation of the results

```
#semFrame1 <-subset(semFrame, (Candidate == "Donald Trump"))
#semFrame2 <-subset(semFrame, (Candidate == "Hillary Clinton"))
#semFrame3 <-subset(semFrame, (Candidate == "Bernie Sanders"))
#fit <-bayes.t.test(semFrame1$score,semFrame2$score)
#show(fit)
#plot(fit)

#fit <- bayes.t.test(semFrame$score)
#plot(fit)

#semFrame$yearF <-factor(semFrame$year, levels = c(2004:2007), labels = c(2004:2007))
#da <-subset(semFrame, select = c(t_term1, yearF))

m0 <-map2stan(alist(score ~ dnorm(mu, sigma),mu <-a
                    ,a ~ dnorm(50, 25),sigma ~ dunif(0.001, 40)),
              data =  semFrame ,iter= 10000, chains = 4, cores = 4 )
```

```
## Computing WAIC
```

```
m1 <-map2stan(alist(score ~ dnorm(mu, sigma),mu <-a[Candidate] ,
                    a[Candidate] ~ dnorm(50, 25),sigma ~ dunif(0.001, 40)),
              data =  semFrame ,iter= 10000, chains = 4, cores = 4 )
```

```
## Computing WAIC
```

```r
compare(m0, m1, func=WAIC)
```

```
##        WAIC       SE    dWAIC      dSE    pWAIC    weight
## m0 1306.134 33.24759 0.000000       NA 2.400944 0.8152923
## m1 1309.103 33.47375 2.969545 2.215846 4.387684 0.1847077
```

```r
precis(m1, depth = 2, prob = .95)
```

```
##            mean         sd       2.5%     97.5%    n_eff     Rhat4
## a[1]  0.4348689 0.11611065 0.20858808 0.6620059 26420.37 0.9999820
## a[2]  0.3041660 0.11393471 0.08352595 0.5285639 26114.87 0.9998937
## a[3]  0.2799999 0.11615614 0.05096678 0.5087748 26171.04 1.0000019
## sigma 1.3085764 0.04700576 1.22052977 1.4037532 24241.79 1.0000037
```

```r
#include your code and output in the document
```

2.1.6.3 Comparison celebrity pair Compare sentiments of celebrity pairs and provide a brief interpretation (e.g. CIs)