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$score, semFrame$Celeb, center = median)
```

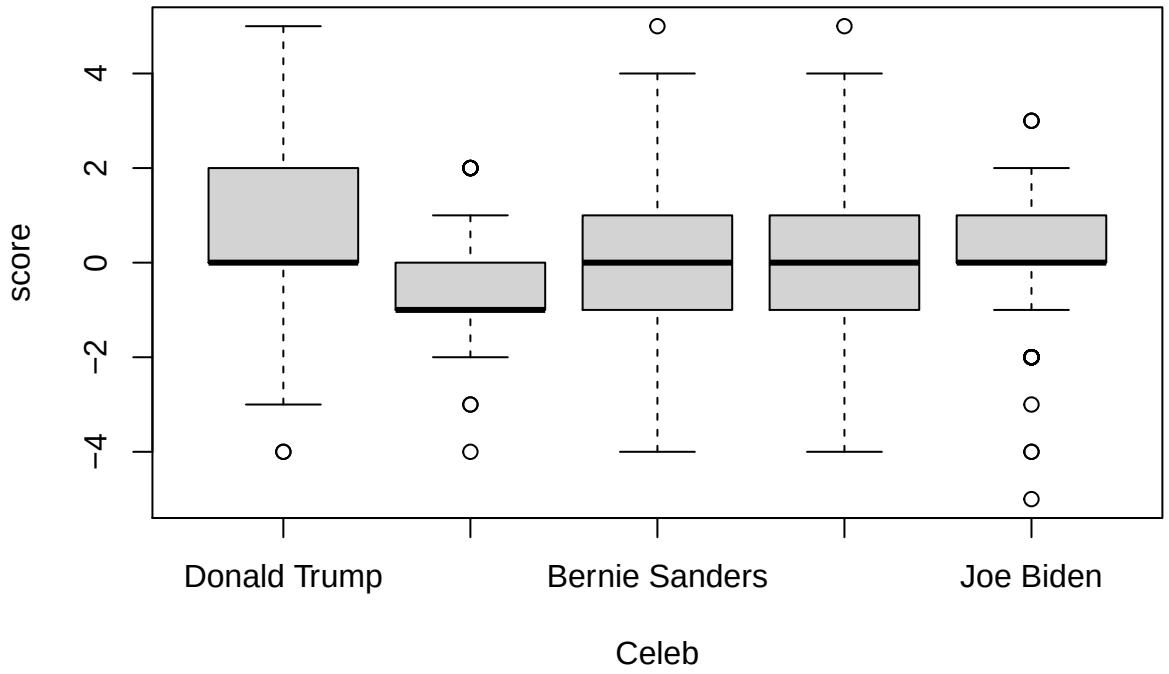
```
## Levene's Test for Homogeneity of Variance (center = median)
##         Df F value    Pr(>F)
## group    4  27.672 < 2.2e-16 ***
##       1495
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The test reveals a p-value smaller than 0.05, indicating that there is significant difference between the group variances in score. We conclude that the variance among the five groups is not equal.

2.1.4 Visual inspection Mean and distribution sentiments

Graphically examine the mean and distribution sentiments of tweets for each celebrity, and provide interpretation We plot both a Boxplot and a distribution histogram which includes a mean line.

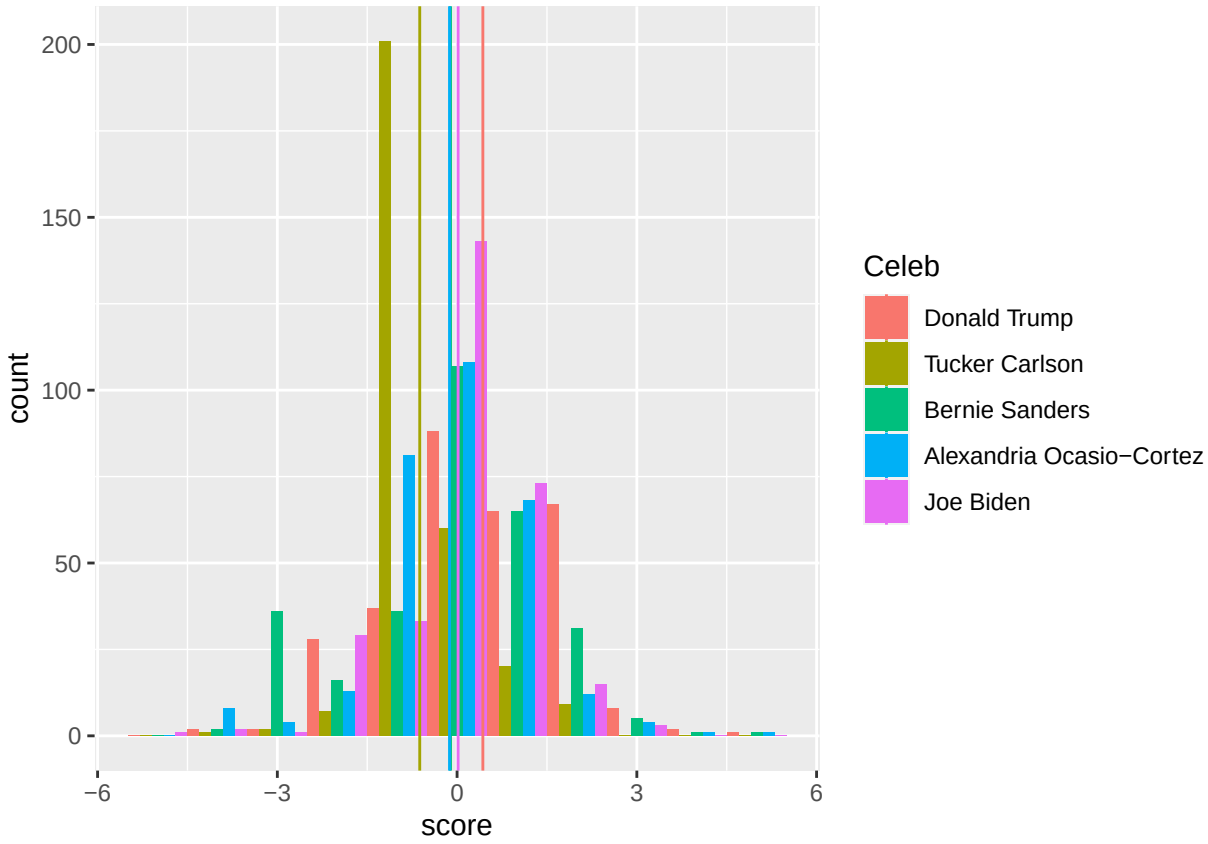
```
boxplot(score ~ Celeb, data = semFrame)
```




```
#sm.density.compare(semFrame$score, semFrame$Celeb, xlab = "Score")
#  title(main="Visual inspection Mean and distribution sentiments")
#  legend('topright', legend = levels(semFrame$Celeb), col=c('red', 'blue', 'green', #'hotpink', 'gold'
#         title="Celebs", lty=1:2, cex=0.8, text.font = 4, bg='lightblue')

cdat <- ddply(semFrame, "Celeb", summarise, score.mean=mean(score))

ggplot(semFrame, aes(x=score, fill=Celeb)) +
  geom_histogram(binwidth=1, position="dodge") +
  geom_vline(data=cdat, aes(xintercept=score.mean,  colour=Celeb),
               linetype="solid", size=0.5)
```




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

We see all US politic Celebs have a mean between -1 and 1. #trump has the highest mean (positive terms)

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 ~ Celeb , data = semFrame)
pander(anova(model0, model1, test = "F"))
```

Table 1: Analysis of Variance Table As the model comparison shows that the predictor improves the fitting.

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1499	2566	NA	NA	NA	NA
1495	2395	4	170.3	26.57	2.554e-21

```
pander(summary(model1))
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.43	0.07308	5.884	4.94e-09
CelebTucker Carlson	-1.053	0.1034	-10.19	1.255e-23
CelebBernie Sanders	-0.54	0.1034	-5.225	1.99e-07

	Estimate	Std. Error	t value	Pr(>|t|)
CelebAlexandria Ocasio-Cortez	-0.5567	0.1034	-5.386	8.355e-08
CelebJoe Biden	-0.4133	0.1034	-3.999	6.666e-05

Table 3: Fitting linear model: score ~ Celeb The p-value of the Celeb variable is low ($p < 0.001$), so it appears that the celeb scraped has a real impact on the sentiment score.

Observations	Residual Std. Error	R^2	Adjusted R^2
1500	1.266	0.06638	0.06388

```
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] 5065.994
```

```
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] 4964.977
```

A lower AIC indicates a better fit, which is the model with the predictor.

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

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

```
## Warning in pander.default(pairwise.t.test(semFrame$score, semFrame$Celeb, : No
## pander.method for "pairwise.htest", reverting to default.
```

- **method**: t tests with pooled SD
- **data.name**: semFrame$score and sem Frame$Celeb
- **p.value**:

Table 4: Table continues below

	Donald Trump	Tucker Carlson	Bernie Sanders
Tucker Carlson	1.255e-22	NA	NA
Bernie Sanders	1.99e-06	7.59e-06	NA
Alexandria Ocasio-Cortez	8.355e-07	1.699e-05	1
Joe Biden	0.0006666	7.645e-09	1

	Alexandria Ocasio-Cortez
Tucker Carlson	NA
Bernie Sanders	NA
Alexandria Ocasio-Cortez	NA
Joe Biden	1

- **p.adjust.method**: bonferroni

We see the pairs of trump-tucker, trump-bernie, trump-aoc, trump-biden differ sigificantly (p<0.001), while tucker-bernie, tucker-biden, bernie-aoc, bernie-biden and aoc-biden dont.

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.

$$score \sim Norm(\mu, \sigma)$$
$$\mu = \alpha$$
$$\alpha = Norm(0, 10)$$
$$\sigma = Uniform(0.001, 10)$$

2.1.6.2 Model comparison

Conduct model analysis and provide brief interpretation of the results

```
m0 <-map2stan(alist(
  score ~ dnorm(mu, sigma),
  mu <-a,
  a ~ dnorm(0, 10),
  sigma ~ dunif(0.001, 10)),
  data =  semFrame ,  iter= 10000, chains = 4,   cores = 4 )
```

```
## Computing WAIC
```

```
m1 <-map2stan(alist(
  score ~ dnorm(mu, sigma),
  mu <-a[Celeb] ,
  a[Celeb] ~ dnorm(0, 10),
  sigma ~ dunif(0.001, 10)),
  data =  semFrame ,iter= 10000, chains = 4, cores = 4 )
```

```
## Computing WAIC
```

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

```
## Warning in class(x) <- NULL: Setting class(x) to NULL; result will no longer be
## an S4 object

## Warning in class(x) <- NULL: Setting class(x) to NULL; result will no longer be
## an S4 object
```

	WAIC	SE	dWAIC	dSE	pWAIC	weight
m1	4971	69.58	0	NA	6.515	1
m0	5066	65.08	94.95	18.92	2.407	2.414e-21

Smaller WAIC is better, so with predictors is the winning model

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

```
## Warning in class(x) <- NULL: Setting class(x) to NULL; result will no longer be
## an S4 object
```

```
## Warning in class(x) <- NULL: Setting class(x) to NULL; result will no longer be
## an S4 object
```

	mean	sd	2.5%	97.5%	n_eff	Rhat4
a[1]	0.4298	0.07327	0.2861	0.5732	34583	0.9999
a[2]	-0.6229	0.07206	-0.7655	-0.4817	33846	1
a[3]	-0.1105	0.07253	-0.2534	0.03179	33776	0.9999
a[4]	-0.1262	0.07264	-0.2696	0.01739	33293	0.9998
a[5]	0.01665	0.07367	-0.1287	0.1604	35785	0.9999
sigma	1.267	0.02287	1.223	1.312	32689	1

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