Report Template coursework assignment A - 2021

CS4125 Seminar Research Methodology for Data Science

Thomas Bos (4543408), Daniël van Gelder (4551028), Jessie van Schijndel (5407397)

20/04/2021

Contents

1 Part 1 - Design and set-up of true experiment **2**

1.1 The motivation for the planned research . . . 2
1.2 The theory underlying the research . . . 2
1.3 Research questions . . . 3
1.4 The related conceptual model . . . 3
1.4.1 Independent Variable (IV) . . . 3
1.4.2 Dependent Variable (DV) . . . 3
1.4.3 Mediating Variable . . . 3
1.4.4 Moderating Variable . . . 3
1.5 Experimental Design . . . 3
1.6 Experimental procedure . . . 4
1.7 Measures . . . 4
1.8 Participants . . . 4
1.9 Suggested statistical analyses . . . 4

2 Part 2 - Generalized linear models **5**

2.1 Question 1 Twitter sentiment analysis (Between groups - single factor) . . . 5
2.1.1 Conceptual model . . . 5
2.1.2 Collecting tweets, and data preparation . . . 5
2.1.3 Homogeneity of variance analysis . . . 5
2.1.4 Visual inspection Mean and distribution sentiments . . . 7
2.1.5 Bayesian Approach . . . 9
2.2 Question 2 - Website visits (between groups - Two factors) . . . 11
2.2.1 Conceptual model . . . 11
2.2.2 Visual inspection . . . 11
2.2.3 Normality check . . . 11

2.2.4 Frequentist Approach 12
2.2.5 Bayesian Approach 12

3 Part 3 - Multilevel model **12**

3.1 Visual inspection 12
3.2 Frequentist approach 12
3.2.1 Multilevel analysis 12
3.2.2 Report section for a scientific publication 13
3.3 Bayesian approach 13
3.3.1 Model description 13
3.3.2 Model comparison 13
3.3.3 Estimates examination 13

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

1.1 The motivation for the planned research

(Max 250 words) The coronavirus pandemic has had a great impact on many aspects of society. University education, in particular, has changed significantly. As education in many countries has shifted from physical lectures to online teleconferencing lectures, concerns have been raised with regards to the effectiveness of this method of education. While the technological developments surrounding teleconferencing have enabled an almost seamless transition from offline to online education, it may be that the lack of a physically present lecturer affects the comprehensibility of the lecture material for students. With this research, we aim to address whether the students' understanding of the lecture material is affected by a different learning setting (i.e., from home watching an online lecture). The results may reveal whether online education is a way to move forward out of the pandemic. Moreover, if the results indicate no significant change in student understanding of material it may open up the way for new form of education, where students could enroll into "digital universities" without needing to be present at any time.

1.2 The theory underlying the research

Figlio et al. (2013) presented, according to them, the first experimental evidence on the effects of live versus online instruction. In this research, participants took an entire microeconomics course either only attending live lectures or online lectures. Exam performance was then compared between both groups and all students which did not volunteer to participate in the experiment but did still follow the course. Result showed that there is a modest difference in exam scores in favour of the students only attending live lectures, although the authors state that the experiments had many limitations and that further research is necessary. In contrary, a more recent survey by Nguyen (2015), which summarizes results of multiple studies, has found that 92% makes online education to be at least as effective, if not better, than live education. However, it is also important to recognize other issues that may arise when switching teaching modalities, which becomes clear when such a shift is forced due to, for example, the onset of COVID-19. In a very recent study by Finnegan (2021), results showed that while results are marginally worse after the shift to online teaching, student experience has deteriorated when their learning environment is suddenly changed, especially with students with poor online access.

1.3 Research questions

Our research question is the following: "How is students' understanding of lecture material affected by attending the lecture live rather than online?". We describe our null hypothesis and alternative hypothesis in the section on suggested statistical analyses.

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)

The following sections describe the conceptual model for each type of variable:

1.4.1 Independent Variable (IV)

The IV of this research is whether the participant (student) attends the lecture physically or from home through online teleconferencing.

1.4.2 Dependent Variable (DV)

The DV of this research is the relative score increase on the test that students make. Before the experiment the participants make a small test regarding the lecture material for which the score is expected to be low as the participants are expected to have no prior knowledge regarding the material. Then after the lecture the students make the same test regarding the lecture material. The relative increase (or unlikely decrease) of score will be the DV.

1.4.3 Mediating Variable

As the students perform the test in a different setting (from home or on campus) depending on the IV. The change in setting is expected to have a mediating effect on the relationship between the IV and DV.

1.4.4 Moderating Variable

There are several factors which may a moderating effect on the relationship between the IV and the DV which are difficult control on the experiment. These mostly have to do with the environment in which the lecture is attended. The following list describes the specific variables which are believed to have this moderating effect:

- (online lecture) video/audio quality
- (online lecture) device that is used to attend lecture (e.g. laptop, tablet, smartphone)
- (both physical and online lecture) presence of noise and/or distraction in environment of watching lecture

1.5 Experimental Design

In order to determine the difference between live and online lectures on students with respect to acquired knowledge the experimental design Pre-test Post-test randomized controlled trail was chosen. This means the participants can be tested before and after the lecture so that the difference in test results, the dependent variable, can be used as an indicator of knowledge gained from said lectures. For the lecture itself, the participants will be divided randomly over live and online groups such that the live group will attend a

lecture face-to-face with a lecturer, and the online group will attend the lecture via an online platform such as Zoom. In order to minimize the influence of moderating variables such as video/audio quality and distractions, the online group will watch the lecture in a quiet, moderated environment on identical systems specifically set up for the experiment.

1.6 Experimental procedure

First, we ask all students in the class who have agreed to participate in our experiment to perform a pre-test a day before the lecture. The pre-test will consist of questions composed by the teacher giving the lecture. The questions should reflect the main learning goals of the lecture. Ideally, this pre-test is done in a controlled setting on campus. If this is not possible due to governmental restrictions, the pre-test is performed online. All students perform the pre-test at the same time. After the pre-test, students are assigned to either the live lecture condition or the online lecture condition. To reduce unexplained variability, we will opt for a randomized block design. We will divide similar participants into blocks based on their pre-test scores. Then, we randomly assign participants from each block to the live condition or the online condition. Students in both conditions will follow the same lecture at the same time. A day after the lecture, the students perform a post-test. Just like the pre-test, the post-test will consist of questions composed by the teacher giving the lecture and should reflect the main learning goals of the lecture. However, the questions from the pre-test should not be repeated. Again, this post-test is ideally done in a controlled setting on campus, but may have to be performed online.

1.7 Measures

In the experiment, both participant groups will take a pre-test and a post-test. This test aims to evaluate the participants' comprehension of the lecture material. The pre-test is meant to serve as a baseline measurement to rule out any pre-existing knowledge of the participants. Both tests will be identical and will be in the form of a multiple choice exam of ten questions to be taken in a short time span (10 minutes). The score of the test is defined as the proportion of correct answers. The measure of the experiment is the ratio between these to tests for each particpant: the score of the post-test divided by the score of the pre-test.

1.8 Participants

Participants should be students and could be recruited by asking for volunteers across a university campus. A small compensation could be offered in return as a sign of appreciation.

1.9 Suggested statistical analyses

First, we determine our null hypothesis H_0 and alternative hypothesis H_1. Our null hypothesis states that there is no difference in student understanding of the lecture material between the two different conditions. Our alternative hypothesis states that there is a difference in student understanding. We create two linear models to predict student understanding of lecture material. First, we create a model which has only an intercept. This model does not use the information about which condition a participant was in. This model will be referred to as m_0. Second, we create a model which does include this information ass a predictor. This model will be referred to as m_1. Then, we compare the fits of the two models to the data. We determine whether m_1 fits significantly better than m_0 through an ANOVA F-test. If this is not the case, we cannot reject our null hypothesis. We may also inspect the significance of the parameters of m_1. If the effect of the condition parameter is not significant, we cannot reject our null hypothesis.

2 Part 2 - Generalized linear models

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

2.1.1 Conceptual model

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.

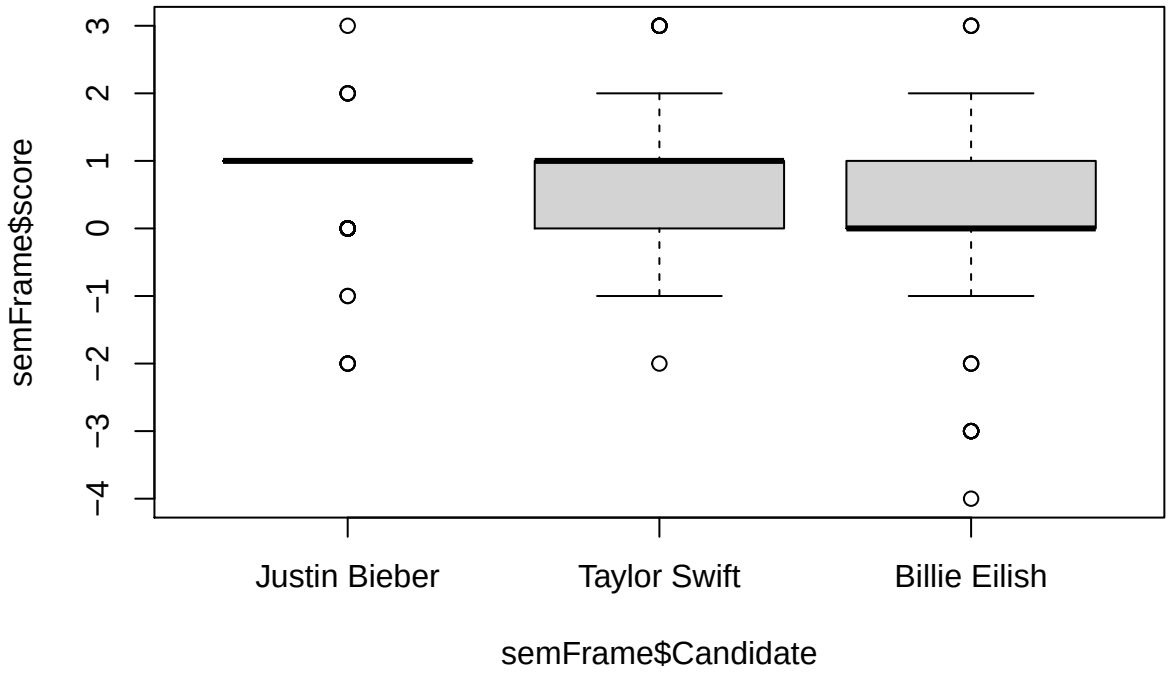
2.1.3 Homogeneity of variance analysis

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

library(car)
```

```
## Loading required package: carData
```

```
boxplot(semFrame$score ~ semFrame$Candidate)
```



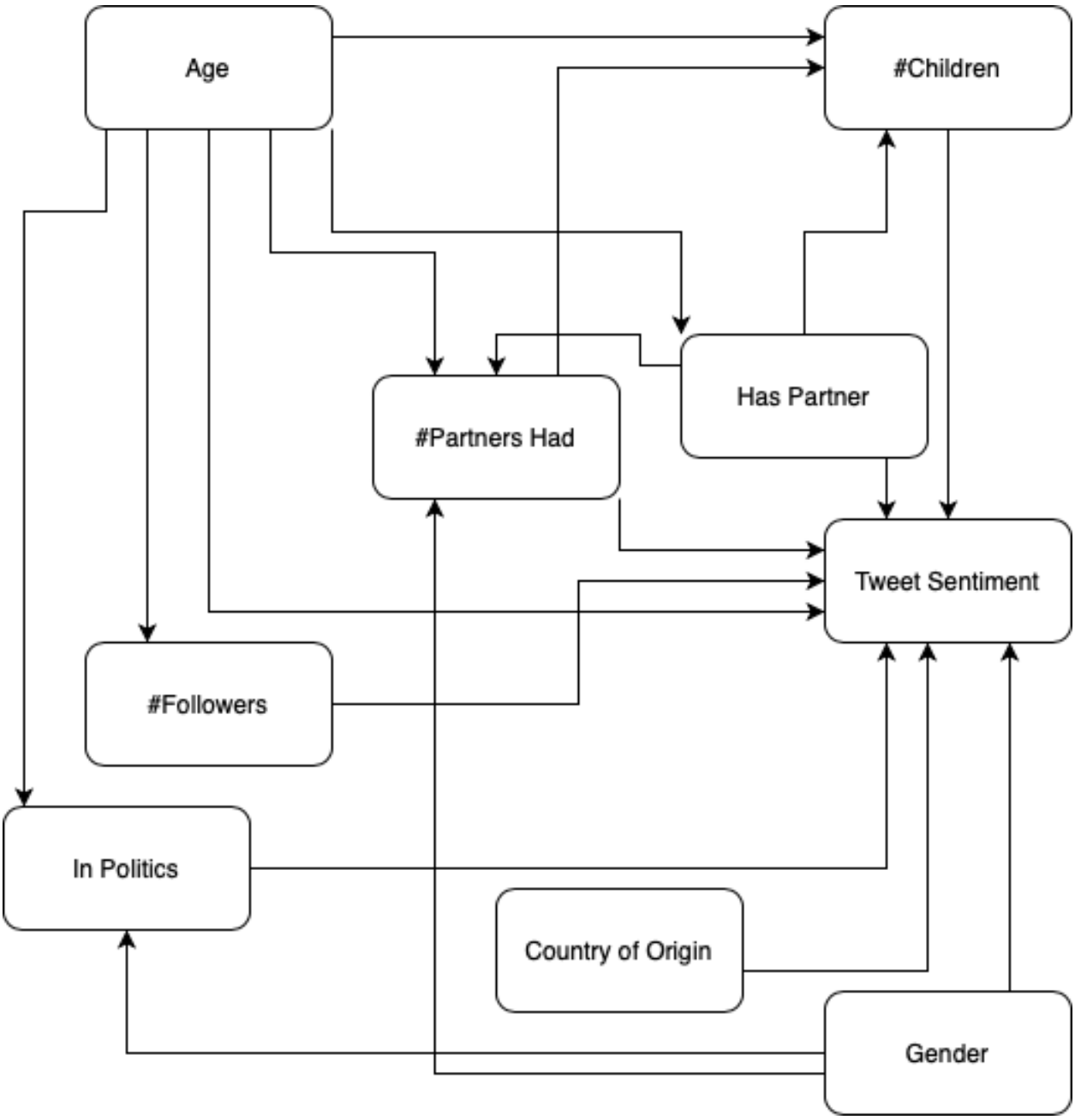


Figure 1: The conceptual model. Different attributes of a celebrity are shown which may influence the sentiment of tweets related to a certain celebrity.

```
leveneTest(semFrame$score, semFrame$Candidate, center = median)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value    Pr(>F)
## group   2  70.347 < 2.2e-16 ***
##       897
## ---
## 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

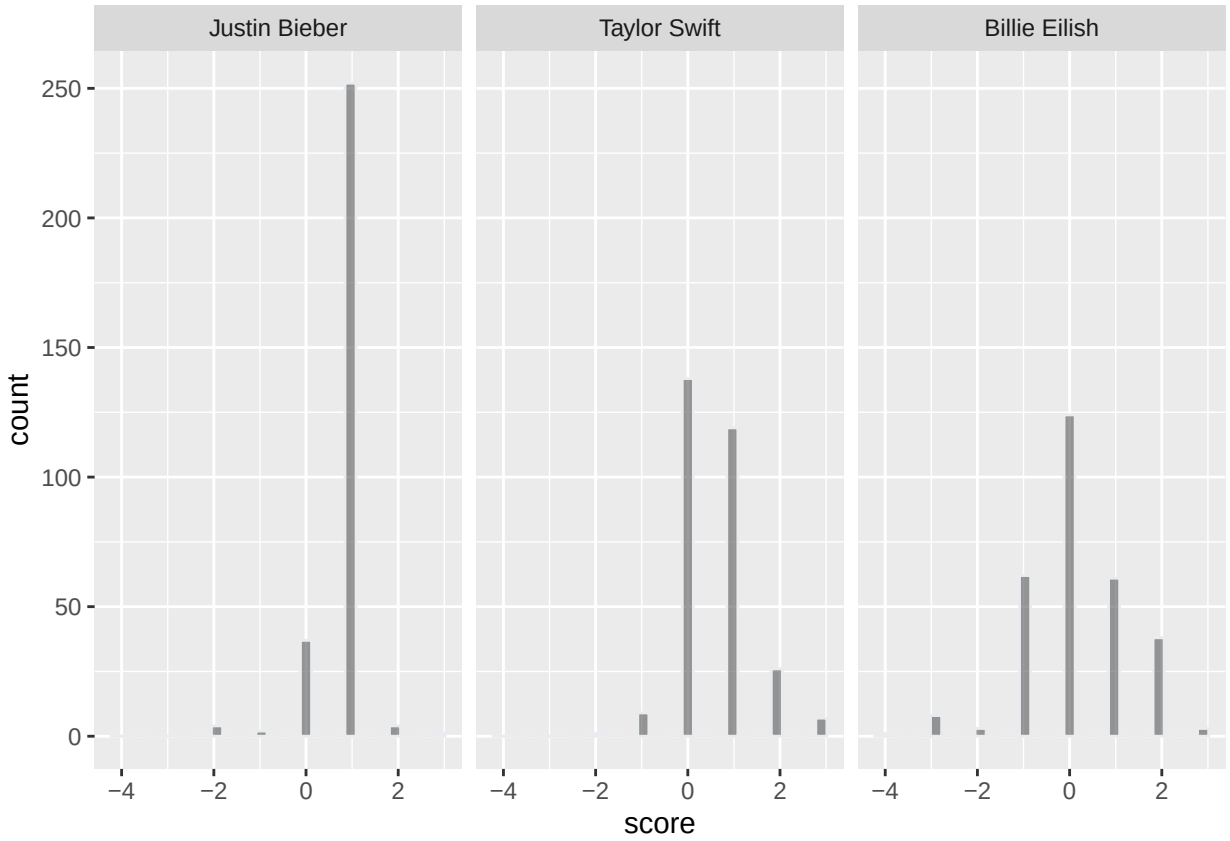
```
#include your code and output in the document

library(ggplot2)
```

```
##
## Attaching package: 'ggplot2'
```

```
## The following object is masked from 'package:NLP':
##
##     annotate
```

```
p <- semFrame %>% ggplot( aes(x=score)) + geom_histogram( color="#e9ecef", alpha=0.6, position = 'identi
plot(p)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Frequentist approach

2.1.4.1 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

```
#include your code and output in the document
library(pander)
library(multcomp)
```

```
## Loading required package: mvtnorm
```

```
## Loading required package: survival
```

```
## Loading required package: TH.data
```

```
## Loading required package: MASS
```

```
##
## Attaching package: 'TH.data'
```

```
## The following object is masked from 'package:MASS':
##
##     geyser
```

```
semFrame$CandidateF <-factor(semFrame$Candidate, levels =c("Justin Bieber", "Taylor Swift", "Billie Eil
#model0 <-lm(score~1, data = semFrame, na.action = na.exclude)
#model1 <-lm(score~CandidateF, data = semFrame, na.action = na.exclude)

res.aov <- aov(score ~ Candidate, data=semFrame)

pander(summary(res.aov))
```

Table 1: Analysis of Variance Model

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Candidate	2	70.62	35.31	47.62	2.215e-20
Residuals	897	665.1	0.7415	NA	NA

2.1.4.2 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(TukeyHSD(res.aov))
```

```
## Warning in pander.default(TukeyHSD(res.aov)): No pander.method for "TukeyHSD",
## reverting to default.No pander.method for "multicomp", reverting to default.
```

- **Candidate**:

	diff	lwr	upr	p adj
Taylor Swift-Justin Bieber	-0.24	-0.4051	-0.07495	0.001935
Billie Eilish-Justin Bieber	-0.6767	-0.8417	-0.5116	0
Billie Eilish-Taylor Swift	-0.4367	-0.6017	-0.2716	2.416e-09

```
#include your code and output in the document
#pairwise.t.test(semFrame$score, semFrame$Candidate, paired = FALSE, p.adjust.method = "bonferroni")
```

2.1.4.3 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.

2.1.5 Bayesian Approach

2.1.5.1 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.

2.1.5.2 Model comparison Conduct model analysis and provide brief interpretation of the results

```
#include your code and output in the document
library("rstan")
```

```
## Loading required package: StanHeaders
```

```
## rstan (Version 2.21.2, GitRev: 2e1f913d3ca3)
```

```
## For execution on a local, multicore CPU with excess RAM we recommend calling
## options(mc.cores = parallel::detectCores()).
## To avoid recompilation of unchanged Stan programs, we recommend calling
## rstan_options(auto_write = TRUE)
```

```
library("rethinking")
```

```
## Loading required package: parallel
```

```
## rethinking (Version 2.13)
```

```
##
## Attaching package: 'rethinking'
```

```
## The following object is masked from 'package:car':
##
##     logit
```

```
## The following object is masked from 'package:stats':
##
##     rstudent
```

```
semFrame <- subset(semFrame, select = c(score, CandidateF))

m0 <-map2stan(
  alist(
    score ~ dnorm(mu, sigma),
    mu <- a,
    a ~ dnorm(0, 2),
    sigma ~ dunif(0.001, 40)),
  data =  semFrame ,iter = 10000, chains = 4, cores = 4
)
```

```
## Computing WAIC
```

```
m1 <-map2stan(
  alist(
    t_term1 ~ dnorm(mu, sigma),
    mu <- a[CandidateF] ,
    a[CandidateF] ~ dnorm(0, 2),
    sigma ~ dunif(0.001, 40)),
  data =  semFrame ,iter = 10000, chains = 4, cores = 4
)
```

```
## Warning: There were 12121 transitions after warmup that exceeded the maximum treedepth. Increase max
## http://mc-stan.org/misc/warnings.html#maximum-treedepth-exceeded

## Warning: Examine the pairs() plot to diagnose sampling problems

## Warning: The largest R-hat is 2.58, indicating chains have not mixed.
## Running the chains for more iterations may help. See
## http://mc-stan.org/misc/warnings.html#r-hat

## Warning: Bulk Effective Samples Size (ESS) is too low, indicating posterior means and medians may be
## Running the chains for more iterations may help. See
## http://mc-stan.org/misc/warnings.html#bulk-ess

## Warning: Tail Effective Samples Size (ESS) is too low, indicating posterior variances and tail quanti
## Running the chains for more iterations may help. See
## http://mc-stan.org/misc/warnings.html#tail-ess

## Computing WAIC

## Error in liks[[j]] : subscript out of bounds
```

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

```
##                mean           sd         2.5%       97.5%     n_eff    Rhat4
## t_term1 0.178168372 8.771176e-01 -1.530732636 1.206046084  2.132556 6.712114
## a[1]    0.178169183 8.771178e-01 -1.530723424 1.206059810  2.132557 6.712094
## a[2]    0.178168566 8.771181e-01 -1.530657234 1.206062269  2.132555 6.712127
## a[3]    0.178167854 8.771174e-01 -1.530760395 1.206082789  2.132556 6.712098
## sigma   0.001001175 1.129388e-06  0.001000057 0.001004196 83.492724 1.010775
```

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

2.2 Question 2 - Website visits (between groups - Two factors)

2.2.1 Conceptual model

Make a conceptual model underlying this research question

2.2.2 Visual inspection

Graphically examine the variation in page visits for different factors levels (e.g. histogram, density plot etc.)

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

2.2.3 Normality check

Visually inspect if variable page visits deviates from a Gaussian distribution, and discuss implication for general linear model analysis.

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

2.2.4 Frequentist Approach

2.2.4.1 Model analysis Conduct a model analysis, to examine the added values of adding 2 factors and interaction between the factors in the model to predict page visits, and include brief interpretation of the results.

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

2.2.4.2 Simple effect analysis If the analysis shows a significant two-way interaction effect, conduct a Simple Effect analysis to explore this interaction effect in more detail.It helps first to look at the means of different conditions in a figure. Provide brief interpretation of the results.

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

2.2.4.3 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.

2.2.5 Bayesian Approach

2.2.5.1 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.

2.2.5.2 Model comparison Conduct model analysis and provide brief interpretation of the results

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

3 Part 3 - Multilevel model

3.1 Visual inspection

Use graphics to inspect the distribution of the score, and relationship between session and score

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

3.2 Frequentist approach

3.2.1 Multilevel analysis

Conduct multilevel analysis and calculate 95% confidence intervals, determine:

- If session has an impact on people score
- If there is significant variance between the participants in their score

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

3.2.2 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.

3.3 Bayesian approach

3.3.1 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.

3.3.2 Model comparison

Select the first 100 participants from the data set. (hint to overcome the Stan problem with a zero index, increase subject id number with 1). Compare models with with increasing complexity.

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

3.3.3 Estimates examination

Examine the estimate of parameters of the model with best fitt, and provide a brief interpretation.

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