Quantitative Methods for Social Science and Business

Final Exam

BSc. International Business and Politics

Copenhagen Business School

Examiner: Zoltan Fazekas

Student number: Date of
submission: 10-06-2019

Pages excluding front page and appendix: 10

Numbers of characters including spaces: 22707

Subquestion 1

The study was performed on Twitter with a sample of 242 treated units, which were sampled based on the last 1000 tweets of 450 randomly chosen users that were at least 6 months old. The treated units were chosen out of certain criteria wherein the primary ones were that the treated samples were white males who had used a certain number of racial slurs targeted at another user determined by an offensiveness score. The 242 samples were assigned to either a control group or one of four treatments, which was to receive a message from a bot telling them that their behaviour was unacceptable within 24 hours of tweeting a racial slur. The bots assigned to the different treatment groups varied in two dimensions: In terms of group identity so that the bots could either be in-group (white) or out-group (black), and in terms of status meaning a low or high number of twitter followers, resulting in four different treatments when combined in all combinations.

The main finding of the article was that subjects who received a treatment from in-group and high-status bots had a significant impact evident with a decrease in the use of racial slurs used by treated subjects. Surprisingly, this only had an impact one-month post-treatment, where after the change in behaviour became insignificant. In addition, it is rather surprising that only one out of the four treatments was effective as none of the other three treatments had any significant impact. Thus, in in this experimental setting, the two dimensions of the treatments are multiplicative rather than additive. The insignificance of the other treatments is surprising, as the treated subjects in all four groups received the same messages about their behaviour and considering that only 3 of the 242 treated subjects accused the bots of being bots, implying that a vast majority of the users acknowledged that the sanction came from a real person and, despite that, they still failed to change their behaviour.

The structure of the experiment with the aspect of randomization aspect provides internal validity for the study research, as it facilitates the ability to access the causal effect of the treatment by comparing the various treated groups with the control groups wherein randomization excludes the possibility of confounding. Furthermore, all the treated groups received the same the message, whereas the only varying factors were race and status of the bots, which further avoided confounding of the results. External validity was maintained by the naturalistic setting on Twitter and the large effort towards making the bots seem real, but it is however difficult to access whether the conclusions can be applied outside of the setting of the study. This is mainly because the experiment was carried out exclusively on twitter towards white males, making it difficult to access if it is applicable in other settings on other races. It is still very relevant today as norms are changing towards a decreased acceptance of racial slur in real-life and the medium for such behaviour, perhaps making this treatment an effective way to combat such hostile behaviour on Twitter or other social media platforms.

The central outcome of the study was that the only the in-group and high-status treatment of the four treatments was effective in accomplishing a decrease in the racist behaviour on twitter. This finding proved the first hypothesis to be true; high-influence and in-group bots have a larger impact. However, the effect of the treatment only lasted for about a month post-treatment and it only worked for one group, indicating that the two dimensions of the treatment were multiplicative. The second hypothesis of the experiment was that anonymous users would be less prone to be affected by the treatment, which was proved wrong as the treatment was not as effective on users with more personal information on their accounts. The dependent variable of the test was carried out by a dictionary method in which the use of the word “n****r” be classified as racist prejudice. An alternative way of measuring harassment on Twitter could be to collect the sample based upon tweets that has been reported as racist. Although this is an equally easy way to collect samples, it would be difficult to achieve the authority needed to acquire this information through Twitter. Collecting a sample through reports and perhaps including a threshold that required a minimum number of reports would exclude the possibility of the use of racial slur in a sarcastic setting, whilst making it far easier to collect a larger sample to examine, hence enabling the experiment to yield results which are more likely to be precise, as a sample of 242 users may not be sufficient to determine the actual effect of the treatment.

Subquestion 2

2.1

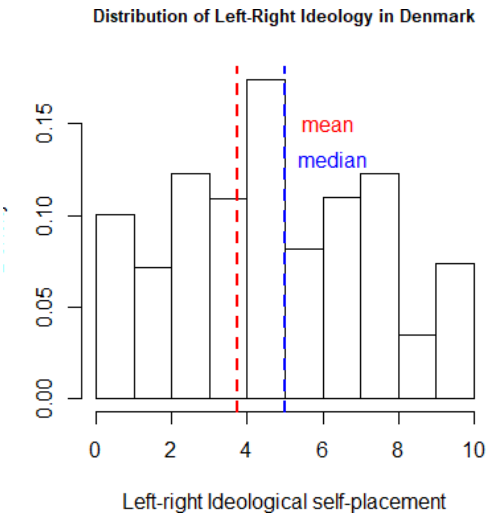
For the t-tests to compare the two countries in terms of political knowledge and religiosity, my null-hypothesis is that there is no difference in the mean between the countries, whereas the alternative hypothesis assumes that there is a difference in the mean. A confidence level of 95% is included in the t-test. For religiosity, the data shows that the mean for self-reported religiosity in the Danish sample is 3.74 whereas it is 6.49 in the Polish sample, which is quite a large difference in the mean. This is also indicated by test-statistic score of -22.368 and further backed up by a very low p-value < 2.2e-16. Therefore, the null hypothesis is rejected, and the alternative hypothesis is accepted meaning that there is a difference between self-reported religiosity in Danish people compared to Polish people.

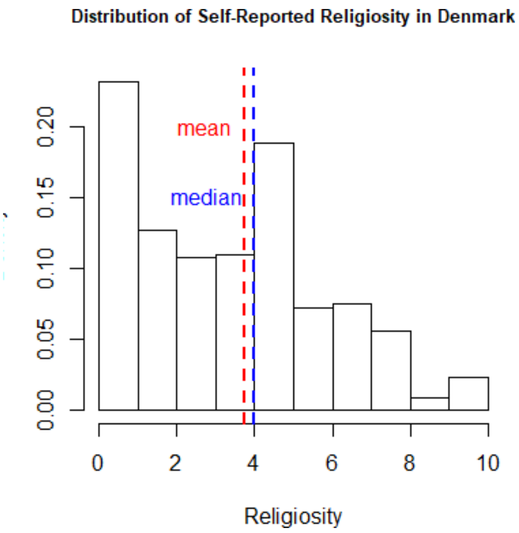
For political knowledge, it is evident that there is a higher degree of agreement, yet the difference in mean is still quite large with the mean for the Danish sample being 4.82 and the mean for the Polish sample being 3.59 showing still quite a large difference between the two populations. This is further backed up with a test statistic score of 14.30 and also a p-value lower than 2.2e-16, therefore also rejecting the null hypothesis for the difference in political knowledge and therefore accepting the alternative hypothesis that there is a difference between the Danish and Polish samples in terms of both self-reported religiosity and political knowledge.

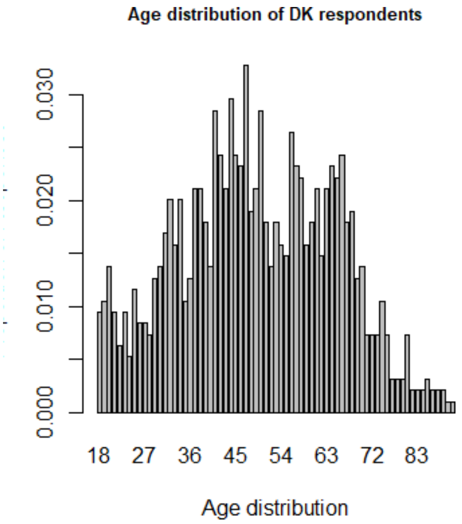
2.2

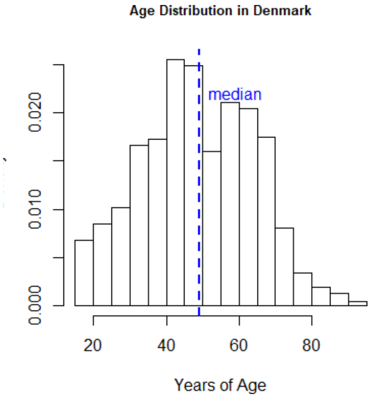
For univariate display of the variables for left-right ideological self-placement, self-reported religiosity and age, I have chosen to visualize the distribution in histograms and a bar plot. The x-axis of the histograms are the various variables, whereas the y-axis is the distribution of responses. The first histogram shown below visualizes the distribution for left-right ideological self-placement (LRSP), wherein it is evident that the distribution is quite evenly distributed around the median of 5, which is also indicated by the mean which is 5.14. Thus, it is indicated by the data that the Danish sample is rather well distributed across the ideological spectrum, with the largest proportion being moderate and a bit higher proportion which is very left-leaning as oppose to right-leaning.

For self-reported religiosity, it is visible on the histogram that there is a stronger tendency for a low self-reported religiosity amongst the Danish people. This is both evident with a mean of 3.74, a median of 4.0, first quartile of 2.0, and a third quartile of 5.0. Interestingly, it is evident that a very large proportion of the Danish sample reports their religiosity as 0 and 1 out of 10 whereas a tiny part of the sample reports themselves as very religious, implying a tendency for the Danish sample to be rather unreligious. The age distribution of the Danish sample is visualized on the third histogram and in the bar plot, wherein it is evident that a range between minimum age of 18 and a maximum age of 93 is represented in the Danish sample. The age distribution is somewhat evenly distributed around the mean of 49.1. However, there is a substantially larger amount of young people relative to the number of subjects above ~70. This is also evident with the mean and the median being 49, indicating that that the difference in age is relatively well distributed but with a large proportion of middle-aged people.









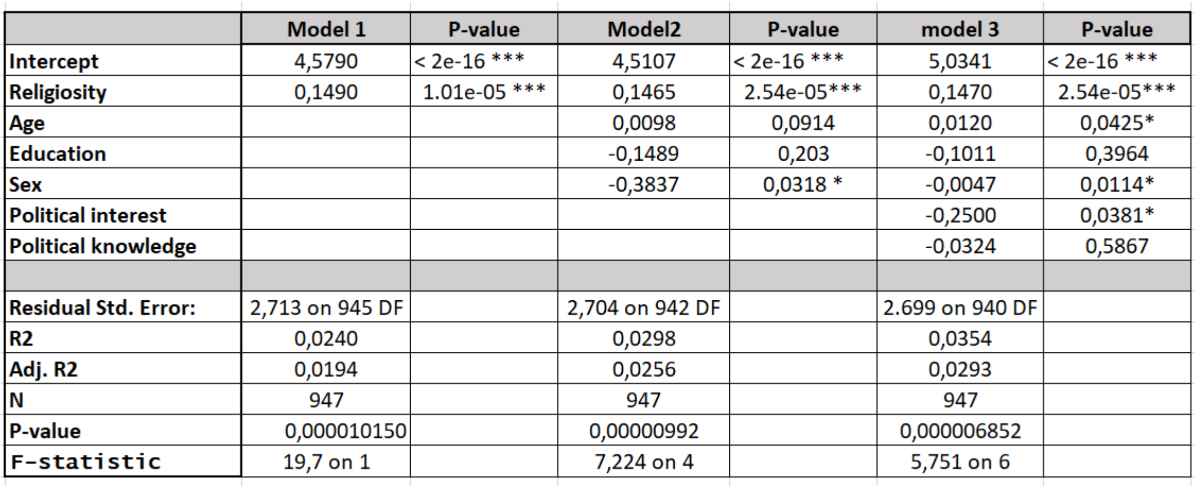
In order to calculate the 95 % confidence interval of the mean of the "LRSP", it is firstly needed to determine the z-score, which with an alpha value of 0.05 is 1.96. Hereafter, I calculated the standard error of the sample for the variable and multiplied that with the z-score to get a margin of error of 0.175. From here, I subtracted and added the margin of error onto the mean of 5.14 to get a confidence interval between the lower bound of 4.96 and 5.31 with a 95% certainty that the true mean for the Danish population is within this interval.

Thus, because the mean of 5.14 of the sample is inside of the 95 % confidence interval, which indicates that that the sample is representative for the Danish population.

The correlation coefficient between self-reported religiosity and left-right ideological self-placement is 0.143 indicating a weak positive relation between the two variables, meaning that the means of the two variables have a weak correlation but not necessarily causation.

2.3 and 2.4

Firstly, three different linear regression models are fitted in R as can be seen in the appendix. Below is the summary of the three different models:



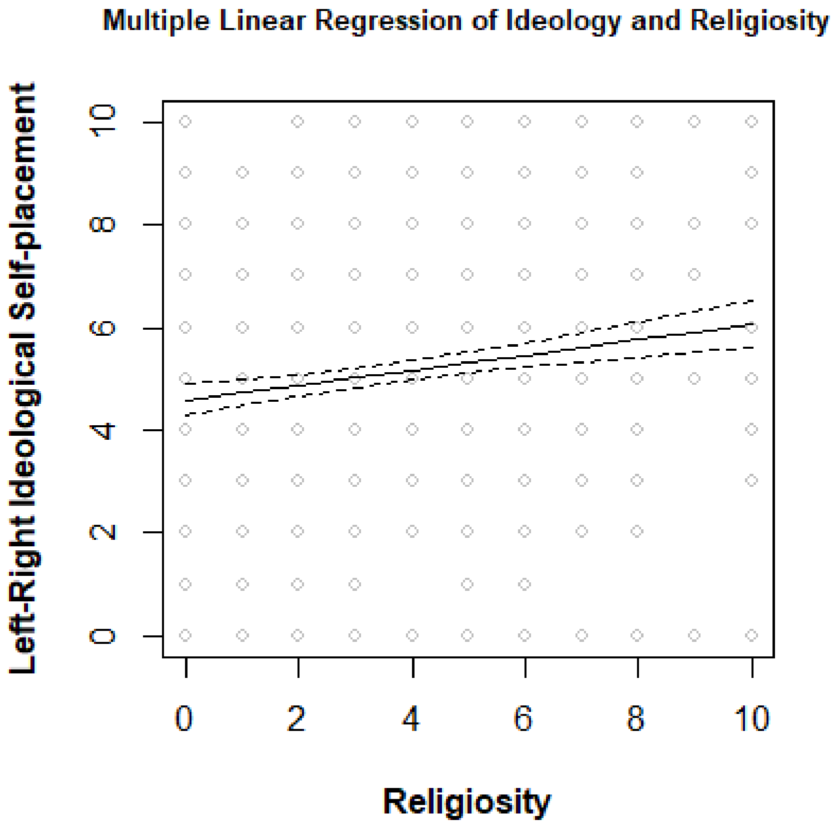
	Model 1	P-value	Model2	P-value	model 3	P-value
Intercept	4,5790	< 2e-16 ***	4,5107	< 2e-16 ***	5,0341	< 2e-16 ***
Religiosity	0,1490	1.01e-05 ***	0,1465	2.54e-05***	0,1470	2.54e-05***
Age			0,0098	0,0914	0,0120	0,0425*
Education			-0,1489	0,203	-0,1011	0,3964
Sex			-0,3837	0,0318 *	-0,0047	0,0114*
Political interest					-0,2500	0,0381*
Political knowledge					-0,0324	0,5867
Residual Std. Error:	2,713 on 945 DF		2,704 on 942 DF		2.699 on 940 DF	
R2	0,0240		0,0298		0,0354	
Adj. R2	0,0194		0,0256		0,0293	
N	947		947		947	
P-value	0,000010150		0,00000992		0,000006852	
F-statistic	19,7 on 1		7,224 on 4		5,751 on 6	

Firstly, it is evident that the estimated intercepts of the three models are quite close to the mean of LPRS which is 5.14. This means that when all the predicting variables are 0, the LRSP is quite close to the mean in all three models but closest in model 3 with all the individual predictors included. It is evident from the coefficients and p-values that religiosity is the most statistically significant in all models, with a p-value very close to 0. The predictors sex, education and political interest are also rather significant with a negative effect on LRSP, implying that a higher level of education, interest in politics and being a female rather than a male makes you more left leaning on average.

The adjusted R-squared-value explains how much of the explanation the model provides when determining LRSP. Thus, the models explain 20%, 26% and 29% of LRSP with all of the predictor variables included. The adjusted R-squared rather than the multiple R-squared is considered in this case, as adding more variables can artificially inflate the R-square value, which the adjusted value penalizes by taking added variables into account. Thus, model 3 is the best of the three models to predict LRSP. The residual standard error on all the models is however approximately 2.7, meaning that on average the models predict LRSP 2.7 wrong with the predictor variables included, which is quite high considering the mean of 5.14. On top of that, the F-score decreases by quite a lot in the different models, meaning that errors in model 2 and 3 have higher influence on the results than in model 1, which makes those models worse in that regard. The p-value of the model does however indicate that all the models are in fact statistically significant in explaining LRSP.

2.5

Below is the visual display of the relationship between self-reported religiosity (x-axis) and LRSP (y-axis), which has all predictors except for religiosity set to their mean value of the sample. As mentioned previously, it is evident that there is on average a weak positive correlation between religiosity and leaning right ideologically. This is also evident in the regression below, with a relatively flat positive slope. This is backed up by the predicted values, as the models predicts an LRSP of 5.32 when religiosity is 5, and an LRSP of 6.06 when religiosity is 10. Lastly, it is evident that there is a higher uncertainty towards the low and especially high values of religiosity, indicating that there is a larger uncertainty amongst the more extreme values, which could be explained by those values representing a smaller proportion of the sample.

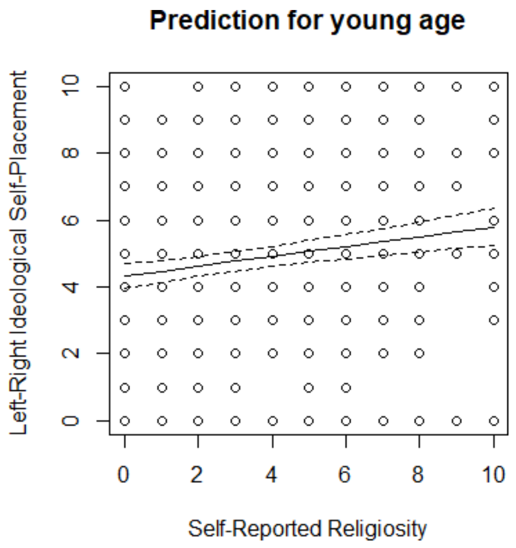


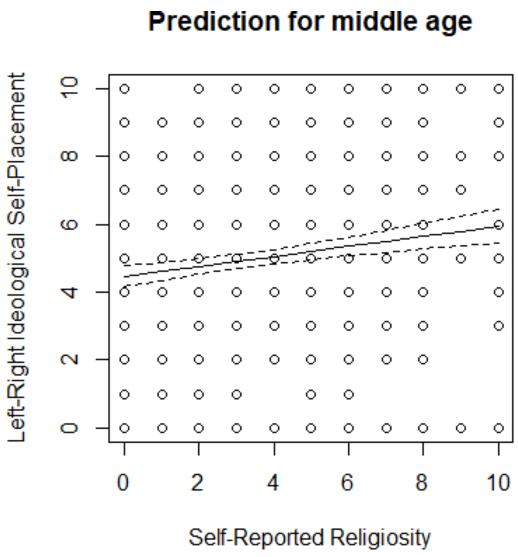
2.6

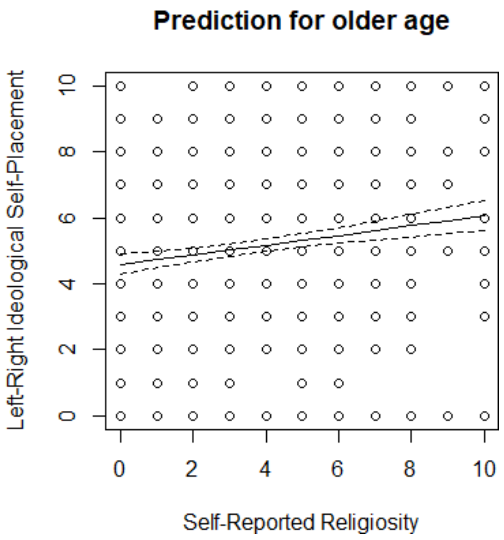
The plots below show that there is not much difference in the patterns between the different age groups selected. The correlation between religiosity and LRSP in the age groups are like the correlation in the previous section. This is also evident when taking the mean of LRSP and religiosity of the first quantile of the age distribution compared to the entire sample. The first quantile of the youngest people has a mean of 3 in religiosity and a mean of 4.78 in LRSP, whereas the mean for the entire sample is 3.74 in religiosity and 5.14 in LRSP, implying a positive correlation between LRSP, religiosity and how old one is, but it does not imply that LRSP and religiosity are conditional on age. Below the predictions for the various ages, the plot shows a prediction of LRSP in relation to age with a level of religiosity of 0 and

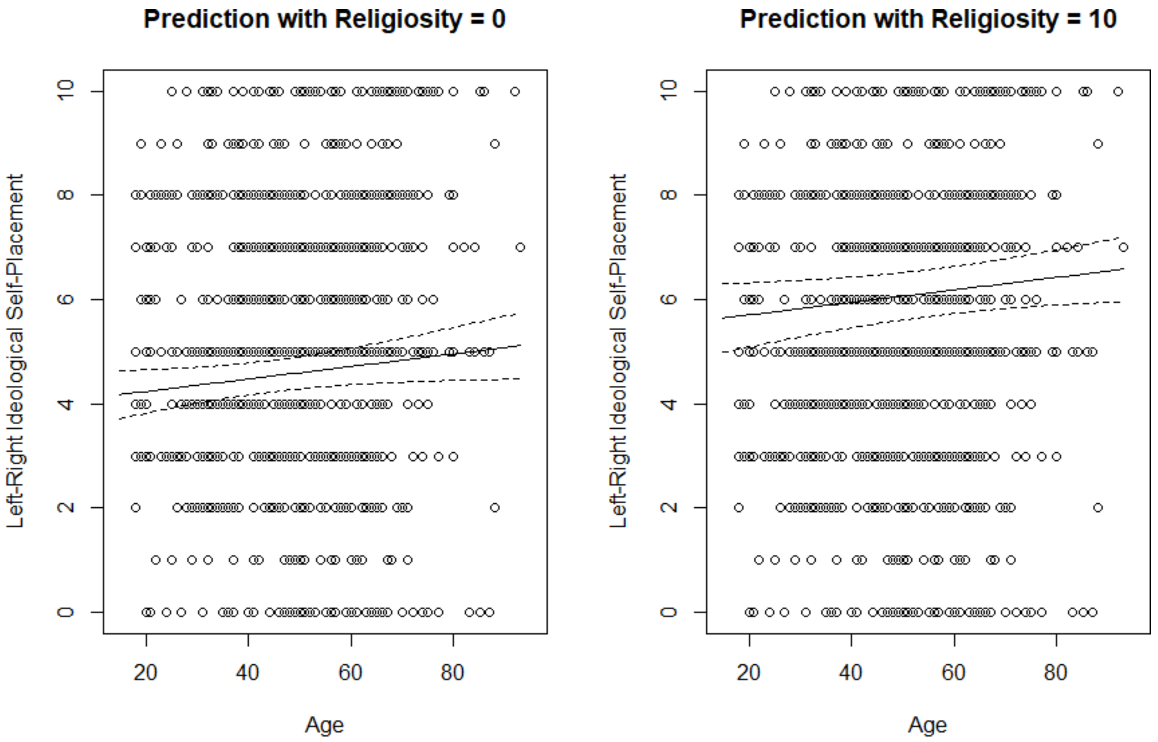
10 respectively. The graphs show quite clearly that religiosity does have a significant impact on LRSP as the slope is approximately the same, but the intercepts differ.

Thus, it can be derived from the data that higher age and religiosity correlates wiith being more right leaning on average. This does however not mean that a high religiosity or a high age causes an individual to be right leaning, but it does imply a advocate possibility for that to be the case. The observational study is limited in terms of determining the relationship, as the sample shows that the Danish population is not very religious. Furthermore, it is limited in cultural and individual factors which might play a significant role towards religiosity and LRSP. An experimental with a more variated design in terms of culture, individuality and other varying societal factors such as average income, gini-coefficent and geographical placement might provide a more accurate answer to the relationship between religiosity and LRSP.









Subquestion 3

The main shortcomings of the articles are that they present some of the data for a finding but leaves out data e.g. presenting how large a proportion of conservatives watch Fox News without explaining why this may be the case, as it might be explained by the greater range of left-leaning media sources compared to right-leaning. The articles also focus quite a bit on media sources and goes rather lightly on social media and when and with whom the respondents discuss politics with, which are also central parts of the study. Furthermore, The Daily Beast article makes many claims from the data that are building on top of the objective data rather than simply presenting it, which could potentially affect the reader's view of the study presented and also it fails to mention anything of the sample size or design of the study, therefore leaving out potentially crucial information about the trustworthiness of the study.

An example of an improvement of the one of those shortcomings is presented below with an improvement of the paragraph from the Daily Beast starting with "For example, 47 percent of …":

"Consistent liberals and conservatives consume entirely different sources of media according to the study, as 47% of the consistent conservatives view Fox News as their main source of information. On the other hand, consistent liberals mainly acquire their information CNN (15%), NPR (13%), MSNBC (12%) etc. On top of that, liberals are generally more trusting in the news sources as they trust in 28 out of 36 whereas conservatives only trust in 8 out of 36 of the news organizations mentioned in the study. This may however be explained with the wider variety of liberal media sources relative to the average web respondent of the study as well as the lack of knowledge towards many of the sources hence making respondents unable to have an opinion on some of the news sources".

The news articles are meant to be relatively short, interesting and easily understandable Therefore, the articles could have used visualizations of the data, which elaborate the different findings of the study in a comprehensive way whilst and includes more data of the related finding, rather than presenting data of the extreme ends of the ideological spectrum hence resulting in a more nuanced presentation of the data. For instance, one visualization of the study shows that only 19% of Fox News' audience are consistently conservative, whereas 37% are mixed in their political views. The articles could have also included more information about section 2 and 3 of the study, wherein the respondents in social media and real-life experiences related to politics are examined, as these sections provide an insight in whether consistent liberals and conservatives seek like-minded individuals. More emphasis on these sections is crucial as they provide explanation of informational filter bubbles, the polarized society and why such phenomena exist.

Subquestion 4

4.1

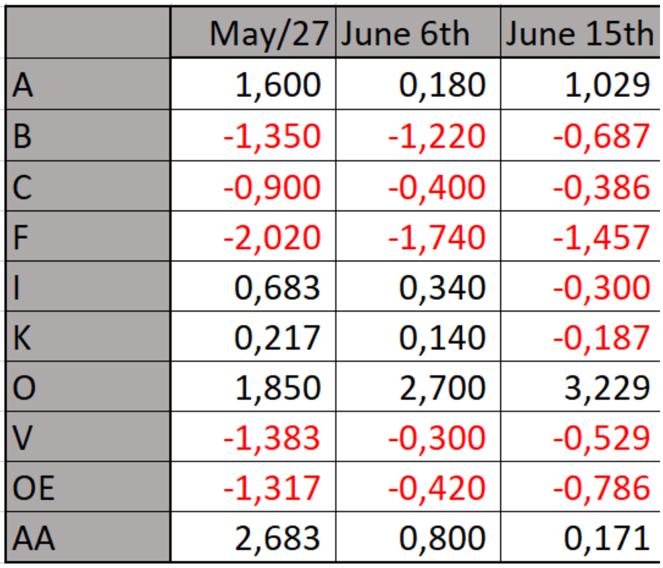
It is assumed that Party K account for the rest of the votes missing from the data of 0.8% so that the total vote count is 100%. The bias and root-mean-squared error (RMSE) are calculated for the various parties. The calculations show that the lowest bias is 0.234 towards Party K, meaning that on average the polls for were most accurate in predicting the results for Party K. The RMSE is the square root of the variance and is lowest for party C with an RMSE of 0.607 meaning that on average the error was lowest for the predictions of this party. Despite the relatively low bias towards Party K and low RMSE of 0.607 for Party C, the bias and RMSE vary quite a lot for the various parties, e.g. Party O for which the bias is 2.889 and the RMSE is 3.070. This suggests that there might be external factors that affect an election, hence making it difficult to predict the actual result before the election.

4.2

Secondly, the bias and RMSE for the polling firms are calculated for the various polling firms. These calculations show that the highest bias of 0.06 registered for Norstat, meaning that the bias for the various firms were generally very low. The highest RMSE is 1.647 for Voxmeter, meaning that they on average had the highest prediction error. Interestingly, the bias and RMSE of the various firms are quite similar compared to the parties, indicating that on average they were all relatively close to the actual result. The RMSE varies from 1.062 for Epinion to the highest RMSE of 1.647 for Voxmeter.

4.3

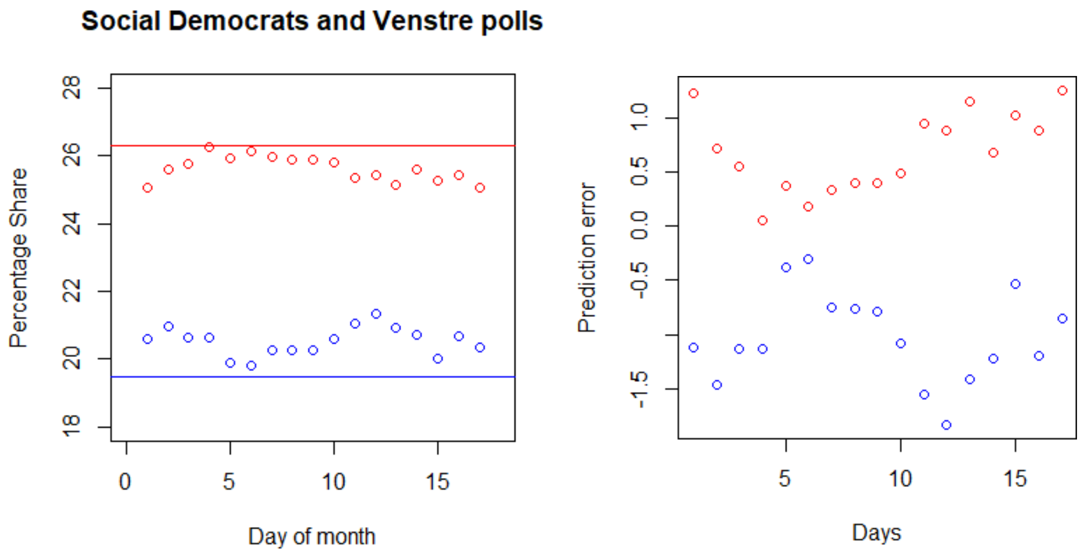
The chosen days are May 27th, June 6th and June 15th. As evident in the table below, some parties are predicted below their actual result whereas others are predicted above throughout the entire period, whereas only party I and K are varying in the campaigning period. The accuracy of the predictions for most of the parties are improved significantly in on the polls of June 15th compared to May 27th except for Party O. Thus, this data indicates that generally polling firms make better predictions when performed closer to the election day.



	May/27	June 6th	June 15th
A	1,600	0,180	1,029
B	-1,350	-1,220	-0,687
C	-0,900	-0,400	-0,386
F	-2,020	-1,740	-1,457
I	0,683	0,340	-0,300
K	0,217	0,140	-0,187
O	1,850	2,700	3,229
V	-1,383	-0,300	-0,529
OE	-1,317	-0,420	-0,786
AA	2,683	0,800	0,171

4.4

The evolution and prediction errors of the two parties Party A (Red) and Party V (blue) from the 1st to the 17th of June are visualized in the two graphs below. First and foremost, it is evident that all the predictions for Party A are below the actual result (The red line), whereas all the predictions for Party V are above the actual result (The blue line). This is evident for the entire period and does not significantly improve towards the election day for the two parties. This is further emphasized on the second graph, wherein it is evident that the prediction errors are approximately between 0 and 1 for Party Ak and 0 and -1.5 for Party V.



Thus, the data show that generally the polls are increasingly accurate towards the election day. However, it was indicated in the table previously and for Party V and Party A in the plot above, the polls tend to either predict the respective parties too low or too high in all polls. Despite that, the polls generally fared quite well in predicting the result taking the many factors of election campaigning as well as the complexity of predicting the outcome of 10 different outcomes into consideration. Furthermore, a limitation of this analysis is the similarity of the parties and the uncertainty of the respondents; there is no category for

“Undecided”, suggesting that respondents who were unsure gave a guess on what they would vote, which could affect the prediction errors drastically.

4.5

The assigned poll is the Megafon poll from 2nd of June 2015 and the parties chosen are Party V and Party C. To calculate the margin of error for the two parties, the Z-score of 1.96 for the confidence level of 95% is first determined. Hereafter, the standard errors for the two parties in the polls are determined and multiplied with the z-score, resulting in a margin of error of 0.0245 for Party V and 0.0103 for Party C. Thus, the margin of error is significantly higher for Party V in relative terms, which might be explained by the size of the parties, as Party V is multiple times larger than Party C in terms of proportion of votes.

4.6

To calculate the margin of error a sample of 2128 rather than 1064, I reassured that the z-score is still 1.96. Hereafter, I calculated a new standard error and multiplied that with the z-score to get the new margin of error of 0.017 for Party V and 0.0073 for Party C. Evidently, the new margin of errors are significantly lower than the previous ones, which can be explained by the larger sample giving an increased precision in predicting the result.

4.7

To calculate what sample size is required for a 1% margin of error with a 99% confidence level, I firstly determined the new z-score of 2.58 with an alfa-value of 0.01. Hereafter, I rearranged the equation and isolated the sample required, as this was now the desired variable to determine. This resulted in a required sample of 11082 for Party V and 1937 for Party C to accomplish a margin of error of 1% with a 99% confidence level.

4.8

I have performed a one-sample test with a confidence level of 95% to conduct the hypothesis-test whether the polling numbers are statistically different. My null-hypothesis was that the predictions of the poll were exactly equal to the observed results at the election. Firstly, I calculated the standard deviation for the sampling distribution of the parties. Secondly, I calculated the upper and lower tails of the distribution in order to get the two-sided p-value of 1.8 for Party V and 0.44 for Party C. Thus, because the poll predictions were 0.211 for Party V and 0.03 for Party C and, additionally, comparing the z-score of -1.27 for Party V and 0.76 for Party C and the critical value of 0.05, we can reject the null-hypothesis hence conclude that the polling numbers are statistically different from the actual values.

Appendix

```
ideo <- read.csv("ideo-ees.csv")
summary(ideo)
str(ideo)
head(ideo)

ideodk <- subset(ideo, syslab == "DK")
ideopl <- subset(ideo, syslab == "PL")

# Firstly, a two-sample t-test is made.
# Ho: Mean difference religiosity in Denmark and Poland are zero.
# Ha: Mean diffrence in religiosity in Denmark and Poland is not equal to zero
#These are 947 respondents from DK and only 758 from Poland

#Performing T-test between religion answers in DK and PL:
ideodk <- subset(ideo, syslab == "DK")
ideopl <- subset(ideo, syslab == "PL")

summary(ideodk)
##Two sample t-test dk religiosity
var(ideodk$relig) #6.90
var(ideopl$relig) #5.90 --> var.eq = F

t.test(ideodk$relig, ideopl$relig, mu=0, alt="two.sided", conf=0.95, var.eq=F, paired = F)



##Two sample t-test for dk political knowledge
var(ideodk$polinfo) #2.55
var(ideopl$polinfo) #3.57 --> var.eq = F

t.test(ideodk$polinfo, ideopl$polinfo, mu=0, alt="two.sided", conf=0.95, var.eq=F, paired = F)
```

2.2

```
###VISUALIZATIONS:
# Visual displays:
x_label <- "Level of Category"
y_label <- "Proportion of Responses"
item_labels <- c("0", "1", "2", "3", "4", "5", "6", "7", "8", "9", "10")
item_labels_age <- c(ideodk$agec[,18:93])



# LRSP disbribution of people in Denmark
hist(ideodk$lrsp, freq = FALSE,
   xlab = "Years of Age", main = "Distribution of Left-Right Ideology in Denmark", cex.main=0.8)
lrspdkmean <- mean(ideodk$lrsp)
abline(v=religdkmean,
    lty = "dashed", lwd = 2.5, col = "red")
lrspdkmedian <- median(ideodk$lrsp)
abline(v = lrspdkmedian,
    lty = "dashed", lwd = 2.5, col = "blue")
text("median", x = lrspdkmedian, y = 0.13, pos = 4, col = "blue")
text("mean", x=lrspdkmean, y = 0.15, pos = 4, col = "red")



###RELIGIOSITY


hist(ideodk$relig, freq = FALSE,
   xlab = "Religiosity", main = "Distribution of self-reported religiosity in Denmark", cex.main=0.8)
religdkmedian <- median(ideodk$relig)
religdkmean <- mean(ideodk$relig)
abline(v=religdkmean,
    lty = "dashed", lwd = 2.5, col = "red")
abline(v = religdkmedian,
    lty = "dashed", lwd = 2.5, col = "blue")
```

```
text("median", x = religdkmedian, y = 0.15, pos = 2, col = "blue")
text("mean", x=religdkmean, y = 0.20, pos = 2, col = "red")


#### AGE
hist(ideodk$age, freq = FALSE,
    xlab = "Years of Age", main = "Age Distribution in Denmark", cex.main=0.8)
ideodk_age_median <- median(ideodk$age)
abline(v = ideodk_age_median,
       lty = "dashed", lwd = 2.5, col = "blue")
text("median", x = ideodk_age_median, y = 0.022, pos = 4, col = "blue")


dk_age_table <- prop.table(table(ideodk$age))
barplot(dk_age_table,
        ylab = y_label, xlab = "Age distribution", main = "Age distribution of DK respondents", names = item_labels_age,
        cex.main=0.8)


#Determine bivariate relationship with correlation
cor(ideodk$lrsp, ideodk$relig)
cor.test(ideodk$lrsp, ideodk$relig, conf.level = 0.95)
boxplot(ideodk$relig, ideodk$lrsp, data=ideodk)
tapply(ideodk$relig, ideodk$lrsp, sd)


#Model 1 for linear regression
#Fitting model 1
model1 <- lm(ideodk$lrsp ~ ideodk$relig, data = ideodk)
model1
summary(model1)


#modél 2 for linear regression
#Fitting model 2
```

```
model2 <- lm(ideodk$lrsp ~ ideodk$relig + ideodk$age + ideodk$educ + ideodk$female, data = ideodk )

model2

summary(model2)


#model 3 for linear regression

#Fitting model 3

model3 <- lm(lrspdk ~ religdk + agedk + educdk + polinfodk + polintdk + sexdk, data = ideodk)

model3


summary(model1)

anova(model1)

par(mfrow = c(2,2))

plot(model1)


summary(model2)

anova(model2)

plot(model2)


summary(model3)

anova(model3)

plot(model3)


mean(lrspdk)


confint(model1,  conf.level  =.95)

confint(model2,   conf.level=.95)

confint(model3, conf.level=0.95)


plot(lrspdk, religdk, main ="scatterplot")

abline(model1)


###
```

```
religdk <- ideodk$relig
lrspdk <- ideodk$lrsp
agedk <- ideodk$age
educdk <- ideodk$educ
polinfodk <- ideodk$polinfo
polintdk <- ideodk$polint
sexdk <- ideodk$female

mreligdk <- mean(ideodk$relig)
mlrspdk <- mean(ideodk$lrsp)
magedk <- mean(ideodk$age)
meducdk <- mean(ideodk$educ)
mpolinfodk <- mean(ideodk$polinfo)
mpolintdk <- mean(ideodk$polint)
msexdk <- mean(ideodk$female)


###Visualization of model 4 Sub.2.7

par(mfrow = c(1,1))
model4 <- data.frame(religdk = seq(from = 0, to = 10, by = 1),
                     agedk = mean(agedk),
                     educdk = mean(educdk),
                     polinfodk = mean(polinfodk),
                     polintdk = mean(polintdk),
                     sexdk = mean(sexdk))


pred <- predict(model3, interval = "confidence",
                level = 0.95,
                newdata = model4)
model4 <- cbind(model4, pred)

plot(religdk, lrspdk,
    xlim = c(0, 10), ylim = c(0, 10),
    xlab = "Religiosity",
    ylab = "Left-Right Ideological Self-placement",
    main = "Multiple Linear Regression of Ideology and Religiosity",
    col = "grey",
    cex = 0.8,
    cex.main = 0.8,
    font.main = 2,
    cex.lab= 1,
    font.lab = 2
)

lines(model4$religdk, model4[, "fit"])
lines(model4$religdk, model4[, "lwr"], lty = "dashed")
lines(model4$religdk, model4[, "upr"], lty = "dashed")



summary(model3)
summary(model4)

model4
coef(model4)
summary(pred)



###Making three different plots with 18-years, 49-years and 93-years respectively:
```

```
par(mfrow = c(1,1))


sortedage <- sort(agedk)
youngideodk <- subset(ideodk, agedk <= quantile(sortedage, 0.25))
medideodk <- subset(ideodk, agedk <= quantile(sortedage, 0.75))
allideodk <- subset(ideodk, agedk <= quantile(sortedage, 1))
mean(youngideodk$relig)
mean(medideodk$relig)
mean(allideodk$relig)

mean(youngideodk$lrsp)
mean(medideodk$lrsp)
mean(allideodk$lrsp)

young <- subset(sortedage, sortedage <= quantile(sortedage, 0.20))

middle <- subset(sortedage, sortedage <= quantile(sortedage, 0.6))

old <- subset(sortedage, sortedage <= quantile(sortedage, 1))

mean(young)
mean(middle)
mean(old)

agepredmin1 <- data.frame(religdk = seq(from=0, to=10, by=1), agedk=mean(young),
               educdk=mean(educdk),sexdk = mean(sexdk),
               polinfodk = mean(polinfodk),polintdk = mean(polintdk))

agepredmin2 <- predict(model3, interval = "confidence", level = 0.95, newdata = agepredmin1)

agepredmin1 <- cbind(agepredmin1, agepredmin2)
```

```
plot(religdk, lrspdk, xlim = c(0, 10), ylim = c(0, 10),
    xlab = "Self-Reported Religiosity",
    ylab = "Left-Right Ideological Self-Placement",
    main = "Prediction for young age")

lines(agepredmin1$religdk, agepredmin1[, "fit"])
lines(agepredmin1$religdk, agepredmin1[, "lwr"], lty = "dashed")
lines(agepredmin1$religdk, agepredmin1[, "upr"], lty = "dashed")

####
median(agedk)

agepredmed1 <- data.frame(religdk = seq(from=0, to=10, by=1), agedk=mean(middle),
                    educdk=mean(educdk),sexdk = mean(sexdk),
                    polinfodk = mean(polinfodk),polintdk = mean(polintdk))

agepredmed2 <- predict(model3, interval = "confidence", level = 0.95, newdata = agepredmed1)

agepredmed1 <- cbind(agepredmed1, agepredmed2)

plot(religdk, lrspdk, xlim = c(0, 10), ylim = c(0, 10),
    xlab = "Self-Reported Religiosity",
    ylab = "Left-Right Ideological Self-Placement",
    main = "Prediction for middle age")

lines(agepredmed1$religdk, agepredmed1[, "fit"])
lines(agepredmed1$religdk, agepredmed1[, "lwr"], lty = "dashed")
lines(agepredmed1$religdk, agepredmed1[, "upr"], lty = "dashed")

######

agepredold1 <- data.frame(religdk = seq(from=0, to=10, by=1), agedk=mean(old),
                    educdk=mean(educdk),sexdk = mean(sexdk),
                    polinfodk = mean(polinfodk),polintdk = mean(polintdk))

agepredold2 <- predict(model3, interval = "confidence", level = 0.95, newdata = agepredold1)

agepredold1 <- cbind(agepredold1, agepredold2)

plot(religdk, lrspdk, xlim = c(0, 10), ylim = c(0, 10),
  xlab = "Self-Reported Religiosity",
  ylab = "Left-Right Ideological Self-Placement",
  main = "Prediction for older age")
lines(agepredold1$religdk, agepredold1[, "fit"])
lines(agepredold1$religdk, agepredold1[, "lwr"], lty = "dashed")
lines(agepredold1$religdk, agepredold1[, "upr"], lty = "dashed")


summary(model4)

#####
#Prediction with religiosity = 0
par(mfrow = c(1,2))
lowrelig <- data.frame(agedk = seq(from=15, to=93, by=1), religdk=0,
                   educdk=mean(educdk),sexdk = mean(sexdk),
                   polinfodk = mean(polinfodk),polintdk = mean(polintdk))


lowrelig2 <- predict(model3, interval = "confidence", level = 0.95, newdata = lowrelig)

lowrelig <- cbind(lowrelig, lowrelig2)

plot(agedk, lrspdk, xlim = c(15, 95), ylim = c(0, 10),
  xlab = "Age",
     ylab = "Left-Right Ideological Self-Placement",
     main = "Prediction with Religiosity = 0")
lines(lowrelig$agedk, lowrelig[, "fit"])
lines(lowrelig$agedk, lowrelig[, "lwr"], lty = "dashed")
lines(lowrelig$agedk, lowrelig[, "upr"], lty = "dashed")

#Prediction with religion = 10:

maxrelig <- data.frame(agedk = seq(from=15, to=93, by=1), religdk=10,
                       educdk=mean(educdk),sexdk = mean(sexdk),
                       polinfodk = mean(polinfodk),polintdk = mean(polintdk))



maxrelig2 <- predict(model3, interval = "confidence", level = 0.95, newdata = maxrelig)

maxrelig <- cbind(maxrelig, maxrelig2)

plot(agedk, lrspdk, xlim = c(15, 95), ylim = c(0, 10),
     xlab = "Age",
     ylab = "Left-Right Ideological Self-Placement",
     main = "Prediction with Religiosity = 10")
lines(maxrelig$agedk, maxrelig[, "fit"])
lines(maxrelig$agedk, maxrelig[, "lwr"], lty = "dashed")
lines(maxrelig$agedk, maxrelig[, "upr"], lty = "dashed")

#Prediction with religion = 0:




#####################################################################################
###

polls <- read.csv("polls-dk.csv")
```

```
polls
#benchmarkpolls <-
summary(polls)
str(polls)
head(polls)


epinion <- subset(polls, pollingfirm == "Epinion")
gallup <- subset(polls, pollingfirm == "Gallup")
greens <- subset(polls, pollingfirm == "Greens")
megafon <- subset(polls, pollingfirm == "Megafon")
norstat <- subset(polls, pollingfirm == "Norstat")
voxmeter <- subset(polls, pollingfirm == "Voxmeter")
wilke <- subset(polls, pollingfirm == "Wilke")
yougov <- subset(polls, pollingfirm == "YouGov")

results_a <- 26.3
results_b <- 4.6
results_c <- 3.4
results_f <- 4.2
results_i <- 7.5
results_o <- 21.1
results_v <- 19.5
results_oe <- 7.8
results_aa <- 4.8
###Thus, assuming that there are not other parties apart from the ones represented in the polls documented,
#party_k received 100-99.2% of the votes, therefore received 0.8 %
results_k <- 0.8


predictedresults <- c(polls$party_a, polls$party_b, polls$party_c, polls$party_f, polls$party_i, polls$partyk, polls$party_o,
                    polls$party_v, polls$party_oe, polls$party_aa)

results <- c(results_a, results_b, results_c, results_f, results_i, results_k, results_o, results_v, results_oe, results_aa)

predicted_epinion <- c(mean(epinion$party_a), mean(epinion$party_b), mean(epinion$party_c), mean(epinion$party_f),
                    mean(epinion$party_i), mean(epinion$party_k), mean(epinion$party_o), mean(epinion$party_v),
                    mean(epinion$party_oe), mean(epinion$party_aa))

predicted_gallup <- c(mean(gallup$party_a), mean(gallup$party_b), mean(gallup$party_c), mean(gallup$party_f),
                    mean(gallup$party_i), mean(gallup$party_k), mean(gallup$party_o), mean(gallup$party_v),
                    mean(gallup$party_oe), mean(gallup$party_aa))

predicted_greens <- c(mean(greens$party_a), mean(greens$party_b), mean(greens$party_c), mean(greens$party_f),
                    mean(greens$party_i), mean(greens$party_k), mean(greens$party_o), mean(greens$party_v),
                    mean(greens$party_oe), mean(greens$party_aa))

predicted_megafon <- c(mean(megafon$party_a), mean(megafon$party_b), mean(megafon$party_c), mean(megafon$party_f),
                    mean(megafon$party_i), mean(megafon$party_k), mean(megafon$party_o), mean(megafon$party_v),
                    mean(megafon$party_oe), mean(megafon$party_aa))

predicted_norstat <- c(mean(norstat$party_a), mean(norstat$party_b), mean(norstat$party_c), mean(norstat$party_f),
                    mean(norstat$party_i), mean(norstat$party_k), mean(norstat$party_o), mean(norstat$party_v),
                    mean(norstat$party_oe), mean(norstat$party_aa))

predicted_voxmeter <- c(mean(voxmeter$party_a), mean(voxmeter$party_b), mean(voxmeter$party_c), mean(voxmeter$party_f),
                    mean(voxmeter$party_i), mean(voxmeter$party_k), mean(voxmeter$party_o),
mean(voxmeter$party_v),

                    mean(voxmeter$party_oe), mean(voxmeter$party_aa))


predicted_wilke <- c(mean(wilke$party_a), mean(wilke$party_b), mean(wilke$party_c),
mean(wilke$party_f),

                  mean(wilke$party_i), mean(wilke$party_k), mean(wilke$party_o),
mean(wilke$party_v),

                  mean(wilke$party_oe), mean(wilke$party_aa))


predicted_yougov <-c(mean(yougov$party_a), mean(yougov$party_b), mean(yougov$party_c),
mean(yougov$party_f),

                  mean(yougov$party_i), mean(yougov$party_k), mean(yougov$party_o),
mean(yougov$party_v),

                  mean(yougov$party_oe), mean(yougov$party_aa))


predicted_a <- mean(polls$party_a)

predicted_b <- mean(polls$party_b)

predicted_c <- mean(polls$party_c)

predicted_f <- mean(polls$party_f)

predicted_k <- mean(polls$party_k)

predicted_i <- mean(polls$party_i)

predicted_o <- mean(polls$party_o)

predicted_v <- mean(polls$party_v)

predicted_oe <- mean(polls$party_oe)

predicted_aa <- mean(polls$party_aa)


#Bias for parties:


#Using package "Metrics" for RMSE and Bias:

library(Metrics)

#Calculated with results - predicted:


bias(results_a, predicted_a) #Bias = 0.788
bias(results_b, predicted_b) #Bias = -0.840
bias(results_c, predicted_c) # Bias = -0.291
bias(results_f, predicted_f) #Bias = -1.657
bias(results_k, predicted_k) # Bias =0.0235
bias(results_i, predicted_i) # Bias = 0.333
bias(results_o, predicted_o) #Bias = 2.889
bias(results_v, predicted_v) #Bias = -1.105
bias(results_oe, predicted_oe) #Bias = -0.803
bias(results_aa, predicted_aa) #Bias = 1.046

#RMSE for parties:

#Calculated by:
rmse_a <- sqrt(mean((results_a - polls$party_a)^2)) #0.788
rmse_a #1.222

rmse(results_a, polls$party_a) #1.222
rmse(results_b, polls$party_b) #1.126
rmse(results_c, polls$party_c) #0.607
rmse(results_f, polls$party_f) #1.785
rmse(results_k, polls$party_k) #0.298
rmse(results_i, polls$party_i) #0.801
rmse(results_o, polls$party_o) #3.070
rmse(results_v, polls$party_v) #1.446
rmse(results_oe, polls$party_oe) #1.130
rmse(results_aa, polls$party_aa) #1.416


#Bias for polling firms:
bias(results, predicted_epinion) #bias = -0.000909
bias(results, predicted_gallup) #bias = 0.01857
bias(results, predicted_greens) # bias = 0.0031818
bias(results, predicted_megafon) # bias = 0.0135

bias(results, predicted_norstat) # bias = 0.06

bias(results, predicted_voxmeter) # vbias = 0.01347826

bias(results, predicted_wilke) #bias = 1.1776682*10^-16 - Lowest bias

bias(results, predicted_yougov) #bias = 0.0075



rmse(results, predicted_epinion) #1.0617

rmse(results, predicted_gallup) #1.267

rmse(results, predicted_greens) #1.165

rmse(results, predicted_megafon) #1.273

rmse(results, predicted_norstat)  #1.513275

rmse(results, predicted_voxmeter) #1.647

rmse(results, predicted_wilke) #1.456

rmse(results, predicted_yougov) #1.083


#######

#Choose one day from each week


may27 <- subset(polls, day == "27", c("party_a", "party_b", "party_c", "party_f","party_k", "party_i",
"party_o",

                              "party_v", "party_oe", "party_aa"))



june6 <- subset(polls, day == "6", c("party_a", "party_b", "party_c", "party_f", "party_k", "party_i",
"party_o",

                             "party_v", "party_oe", "party_aa"))


june15 <- subset(polls, day == "15", c("party_a", "party_b", "party_c", "party_f", "party_k",
"party_i", "party_o",

                              "party_v", "party_oe", "party_aa")
```

```
#Calculate the prediction error for each party for each of the days:
results_a - mean(may27$party_a) #1.6
results_b - mean(may27$party_b) #-1.35
results_c - mean(may27$party_c) #-0.9
results_f - mean(may27$party_f) #-2.02
results_i - mean(may27$party_i) #0.683
results_k - mean(may27$party_k) #0.217
results_o - mean(may27$party_o) #1.85
results_v - mean(may27$party_v) #-1.383
results_oe - mean(may27$party_oe) #-1.317
results_aa - mean(may27$party_aa) #2.683

results_a - mean(june6$party_a) #0.18
results_b - mean(june6$party_b) #-1.22
results_c - mean(june6$party_c) #-0.4
results_f - mean(june6$party_f) #-1.74
results_i - mean(june6$party_i) #0.34
results_k - mean(june6$party_k) #0.14
results_o - mean(june6$party_o) #2.7
results_v - mean(june6$party_v) #-0.3
results_oe - mean(june6$party_oe) #-0.42
results_aa - mean(june6$party_aa) #0.8

results_a - mean(june15$party_a) #1.029
results_b - mean(june15$party_b) #-0.6867
results_c - mean(june15$party_c) #-0.386
results_f - mean(june15$party_f) # -1.457
results_i - mean(june15$party_i) # -0.3
results_k - mean(june15$party_k) # -0.186
results_o - mean(june15$party_o) # 3.229
results_v - mean(june15$party_v) #-0.529
results_oe - mean(june15$party_oe) # -0.786
```

```
results_aa - mean(june15$party_aa) # 0.171


####Create a plot

Day1 <- subset(polls, day == "1")
Day2 <- subset(polls, day == "2")
Day3 <- subset(polls, day == "3")
Day4 <- subset(polls, day == "4")
Day5 <- subset(polls, day == "5")
Day6 <- subset(polls, day == "6")
Day7 <- subset(polls, day == "7")
Day8 <- subset(polls, day == "8")
Day9 <- subset(polls, day == "9")
Day10 <- subset(polls, day == "10")
Day11 <- subset(polls, day == "11")
Day12 <- subset(polls, day == "12")
Day13 <- subset(polls, day == "13")
Day14 <- subset(polls, day == "14")
Day15 <- subset(polls, day == "15")
Day16 <- subset(polls, day == "16")
Day17 <- subset(polls, day == "17")

Days1 <- 1
Days2 <- 2
Days3 <- 3
Days4 <- 4
Days5 <- 5
Days6 <- 6
Days7 <- 7
Days8 <- 8
Days9 <- 9
```

```
Days10 <- 10
Days11 <- 11
Days12 <- 12
Days13 <- 13
Days14 <- 14
Days15 <- 15
Days16 <- 16
Days17 <- 17


social1 <- mean(Day1$party_a)
social2 <- mean(Day2$party_a)
social3 <- mean(Day3$party_a)
social4 <- mean(Day4$party_a)
social5 <- mean(Day5$party_a)
social6 <- mean(Day6$party_a)
social7 <- mean(Day7$party_a)
social8 <- mean(Day8$party_a)
social9 <- mean(Day9$party_a)
social10 <- mean(Day10$party_a)
social11 <- mean(Day11$party_a)
social12 <- mean(Day12$party_a)
social13 <- mean(Day13$party_a)
social14 <- mean(Day14$party_a)
social15 <- mean(Day15$party_a)
social16 <- mean(Day16$party_a)
social17 <- mean(Day17$party_a)


venstre1 <- mean(Day1$party_v)
venstre2 <- mean(Day2$party_v)
venstre3 <- mean(Day3$party_v)
```

```
venstre4 <- mean(Day4$party_v)
venstre5 <- mean(Day5$party_v)
venstre6 <- mean(Day6$party_v)
venstre7 <- mean(Day7$party_v)
venstre8 <- mean(Day8$party_v)
venstre9 <- mean(Day9$party_v)
venstre10 <- mean(Day10$party_v)
venstre11 <- mean(Day11$party_v)
venstre12 <- mean(Day12$party_v)
venstre13 <- mean(Day13$party_v)
venstre14 <- mean(Day14$party_v)
venstre15 <- mean(Day15$party_v)
venstre16 <- mean(Day16$party_v)
venstre17 <- mean(Day17$party_v)


socialvenstredays <- c(social1, venstre1, social2, venstre2, social3, venstre3, social4, venstre4,
social5, venstre5,
                       social6, venstre6, social7, venstre7, social8, venstre8, social9, venstre9, social10,
venstre10,
                       social11, venstre11, social12, venstre12, social13, venstre13, social14, venstre14,
social15, venstre15,
                       social16, venstre16, social17, venstre17)
dayss <- c(Days1,Days1, Days2, Days2, Days3, Days3, Days4, Days4, Days5, Days5, Days6, Days6,
Days7,
           Days7, Days8, Days8, Days9, Days9, Days10, Days10, Days11, Days11, Days12, Days12,
Days13,
           Days13, Days14, Days14, Days15, Days15, Days16, Days16, Days17, Days17)


plot(dayss, socialvenstredays,
     xlab="Day of month",
     ylab="Percentage Share",
     main="Social Democrats and Venstre polls", col=ifelse(socialvenstredays>=23, "red", "blue"),
xlim=c(0, 18), ylim = c(18,28))
abline(a=26.3, b=0, col="red")
abline(a=19.5, b=0, col="blue")
results_v
results_a


preds1 <- results_a - social1
preds2 <- results_a - social2
preds3 <- results_a - social3
preds4 <- results_a - social4
preds5 <- results_a - social5
preds6 <- results_a - social6
preds7 <- results_a - social7
preds8 <- results_a - social8
preds9 <- results_a - social9
preds10 <- results_a - social10
preds11 <- results_a - social11
preds12 <- results_a - social12
preds13 <- results_a - social13
preds14 <- results_a - social14
preds15 <- results_a - social15
preds16 <- results_a - social16
preds17 <- results_a - social17

predv1 <- results_v - venstre1
predv2 <- results_v - venstre2
predv3 <- results_v - venstre3
predv4 <- results_v - venstre4
predv5 <- results_v - venstre5
predv6 <- results_v - venstre6
predv7 <- results_v - venstre7
predv8 <- results_v - venstre8
predv9 <- results_v - venstre9
```

```
predv10 <- results_v - venstre10

predv11 <- results_v - venstre11

predv12 <- results_v - venstre12

predv13 <- results_v - venstre13

predv14 <- results_v - venstre14

predv15 <- results_v - venstre15

predv16 <- results_v - venstre16

predv17 <- results_v - venstre17


predictionerrors <- c(preds1, predv1, preds2, predv2, preds3, predv3, preds4, predv4, preds5, predv5,
preds6, predv6,
                 preds7, predv7, preds8, predv8, preds9, predv9, preds10, predv10, preds11, predv11,
preds12, predv12, preds13, predv13,
                 preds14, predv14, preds15, predv15, preds16, predv16, preds17, predv17)


plot(dayss, predictionerrors,
    xlab = "Days",
    ylab = "Prediction error", col=ifelse(predictionerrors>=0, "red", "blue"))



### Select random poll
poll_unique <- unique(polls$poll)
set.seed(128956)


poll_selected <- sample(poll_unique, 1)
poll_selected #[1] Megafon-2015-06-02


megafonpoll <- subset(polls, id=="762")
summary(megafonpoll)
# Largest parties: Social Demokratiet = 25.3, Venstre = 21.1, Dansk Folkeparty = 18.2
# Smallest parties: Kristendemokraterne = 1.3, Konservativ Folkeparti = 3, Alternativet = 3.3


# Parties chosen are Venstre (party_v) and konservativ folkeparti (party_c)
```

```
summary(megafonpoll)
#S
# Firstly, follwing are the mean of the parties, the results, and the results of my assigned poll_


meanc <- mean(polls$party_c) #mean is 3.69
meanv <- mean(polls$party_v) #20.6053
meanv
results_c #3.4
results_v #19.5
megafonpoll$party_c # 3
megafonpoll$party_v #21.1


#Calculating Z-score with 95 % confidence interval
m <- megafonpoll$n
m #1064 is the sample size

conf.level <- 0.95
z1_score <- qt(((1+conf.level)/2), df=m-1)
z1_score #z-score = 1.96
conf.level <- 0.95

#Thus, the z-score for both party v and party c with a confidence interval of 95% and a sample size of 132 is 1.96

#Party_v margin error:
#standard error:
megafonpoll$party_v #0.211

marg_v <- (z1_score*(sqrt((0.211*(1-0.211))/m)))
marg_v #margin of error for Venstre is 0.02454436
```

```
#Party_c margin of error:
megafonpoll$party_c #0.03
marg_c <- (z1_score*(sqrt((0.03*(1-0.03))/m)))
marg_c #0.01026168



#Now we are doubling the sample, therefore it is m*2
2*m # Sample = 2128
#Firstly, we calculate the margin of error with the doubled sample for Venstre:
z2_score <- qt(((1+conf.level)/2), df=(2*m)-1)
z2_score #z-score = 1.96



marg_v2 <- (z2_score*(sqrt((0.211*(1-0.211))/2128)))
marg_v2 #0.01735 = 1.735 %
#Then I calcualate the margin of error with the doubled sample for Konservative:
marg_c2 <- (z2_score*(sqrt((0.03*(1-0.03))/2128)))
marg_c2 # 0.00725 = 0.73%

#Now, we are trying to figure out what sample size is required to have a margin of error
#of one percent, when the confidence interval is 99%. With a confidence interval of 99%, the new z-score is 2.58:


samplereq_v <- ((0.211*(1-0.211)*((2.58/0.01)^2)))
samplereq_v #11081.51 is the required sample size for party v


samplereq_c  <- ((0.03*(1-0.03)*((2.58/0.01)^2)))
samplereq_c # 1937.012 is the required sample size for party c



###
#The Null Hypothesis: H0 = The polling prediction for Venstre is equal to 0.211 (21.1%)
#Thus, the null value is 0.211
```

```
#Ha: (Alternative hypothesis): polling prediction for Venstre is NOT equal to 0.211 (21.1%)
#Observed value (Results for Venstre) (results_v1) = 0.195 (19.5%)
#Confidence level is 95% (alpha value = 0.05) and the sample size is 1064 on the megafon poll.
m <- 1064
results_v
megafonpoll$party_v
results_v1 <- 0.195
sdparty_v <- sqrt((0.211*(1-0.211))/m)
lowerv <- pnorm(0.211 - (results_v1 - 0.211), mean = 0.211, sd=sdparty_v)
upperv <- pnorm(results_v1, mean=0.211, sd=sdparty_v, lower.tail = FALSE)
twosidedpvalueo <- upperv + lowerv

zscorev <- (results_v1 - 0.211)/sdparty_v
2*pnorm(zscoreo, lower.tail = FALSE)

#pvalue is less than Null, thus we reject the null hypothesis, which means the prediction was not the actual results
#meaning the prediction was off
#Thus the percentage difference I observed was unlikely to appear if left to chance alone
#Statistically significent

#The Null Hypothesis: H0 = The polling prediction for Konservative Folkeparti is equal to 0.03 (3%)
#Thus, the null value is 0.03
#Ha: (Alternative hypothesis): polling prediction for Venstre is NOT equal to 0.03 (3%) meaning that the 1
#Observed value (Results for Konservative) (results_c1) = 0.034 (3.4%)
#Confidence level is 95% (alpha value = 0.05) and the sample size is 1064 on the megafon poll.

m <- 1064
results_c
megafonpoll$party_c
results_c1 <- 0.034
sdpartyc <- sqrt((0.03*(1-0.03))/m)
```

```
lowerc <- pnorm(0.03 - (results_c1 - 0.03), mean = 0.03, sd=sdpartyc)
upperc <- pnorm(results_c1, mean=0.03, sd=sdpartyc, lower.tail = FALSE)
twosidedpvaluec <- (upperc + lowerc)

zscorec <- (results_c1 - 0.03)/sdpartyc
2*pnorm(zscorec, lower.tail = FALSE)
```