Quantitative methods for Business and Social Science

Final Exam

BSc in International Business and Politics

Copenhagen Business School

Pages: 10,9 (13 including frontpage & bibliography)

Characters: 24.818

Reference system: APA 7th edition

Supervisor: Zoltan Fazekas

01. May

Question 1.1

The two experiments are key components of a True Experimental Design (Staller, 2010) which must include a control group, a manipulatable variable and most importantly the assignment of the treatment variable is done randomly, creating theoretical identical treated and control groups. The two experiments are needed to test the proposed hypotheses of causality between the individuals' rating of USPS and their implicit attitudes towards the USPS. Because both experiments are set up as True Experimental Designs, they utilize randomized controlled trials (RCTs). This means that in both of the experiments there is an equal chance of being in the control group or in the treatment group. This is the strongest way of isolating the effect of the treatment variable, and therefore, the preferred benchmark of determining causality. The test setup therefore has a high internal validity which contributes to the causal assumptions. Because these two experiments are designed through RCTs they can overcome the challenges of causal inference and its counter factual outcomes, by approximating an average treatment effect, thereby helping us to draw possible causal conclusions.

The two experiments differ in all three stages, in some substantive ways. Firstly, even though both experiments utilize an IAT test, Experiment 1 focuses the test persons stereotypical presumption of the USPS and FedEx, distinguishing between *fast* and *slow*. However, in the second experiment a preference focus is applied, testing people's association of *good (FedEx)* and *bad (USPS)*. Secondly, Experiment 2 consists of three groups, including an "advertising" treatment variable which is not included in the first experiment, that only consists of two groups. In the third stage, Experiment 1 includes a longitudinal aspect, randomly assigning some test persons to do their USPS rating immediately and some after a 2 day delay. This aspect adds another layer of data, enabling researchers to evaluate the effect of the treatment variables over time, i.e., if positive USPS information or advertisement show an effect, then is it a lasting effect or short lived. This longitudinal aspect is not present in the second experiment. Because the second experiment utilizes a preference IAT instead of a stereotype IAT, it is possible to isolate a potential bias of the test persons as they are known to belong predominately to a social class with a more positive attitude towards the public sector.

Question 1.2 (a)

The Implicit Attitude Test (IAT) is a test meant to quantify association between two concept categories and two attribute categories. Meaning that researchers can measure what level of association there is between specific feelings, attitudes etc. and specific institutions, objects, etc. Simplified, this is done by coupling an attributive category to a concept category, measuring the time needed so sort a list of items into the two different concept categories and their attributive category. Thereafter, the concept categories change attributive categories, and the sorting task is again timed. The times are compared, and this often results in a quicker time for sorting the items, where the concept category was matched with the test persons ideas of the appropriate attributive category.

The IAT test score is an interval variable, more specifically a continuous variable, which can take infinitely many values, as it is composed of two measures of time subtracted from each other. The researchers in this paper, argues that "individuals' implicit attitudes regarding public sector organizations are biased". Therefore, the researchers utilize the IAT test as a way to translate and quantify abstract concepts as "implicit biases" into data that can be used for further investigation. By using the IAT test instead of a survey for measurement the researchers avoid potential unit and item nonresponses, which can be problematic if occurrence is nonrandom, they also overcome issues of acquiescence biases and satisficing bias. Furthermore, because of the practicality of the test, especially the speed at which association sorting happens and the anonymity of answering alone on a computer, the problem of social desirability biases is eliminated.

Question 1.2 (b)

The internal validity of the study seems strong, as the methodology shows high control, and all experiments undergo RCTs. Furthermore, the choice of measurement (IATs) eliminates multiple possible biases which further supports the validity of the study. However, as the author describes the data collection process through MTurk, a sample selection bias of non-random selection must be considered. This is proven as the data samples are predominantly educated, liberal white males, which lowers its external validity. Nevertheless, the author argues that the data sample has a high level of Inter-rater Reliability (Lange, 2011), arguing that various other scholars have reached empirical success with MTurk samples. This could reduce some issues regarding the lack of external validity of the samples.

Question 1.3

The results of the main effects of Experiment II presented in table 3 are the IAT control group, the information- and the advertising variable. Firstly, by looking at the -0,20 coefficient for the IAT control group we see that a 1 unit increase in the IAT test score on average is associated with a 0,20 unit decrease of USPS performance rating. However, as this coefficient is linked with a p-value of 0.551, the results must be classified as statistical insignificant as the study follows the conventional P<0,05, requirement for statistical significance. Nevertheless, the uncertainty of the experiment is lowered when looking at the standard error (SE). The IAT score coefficient has a relatively low SE score of 0.34, suggesting that even though the sample is nonrandom it relatively strong represents the true population. The researchers put forward a hypothesis H1, which states that individuals' implicit attitudes affect their performance rating of the USPS. This hypothesis is generally supported by the data in Experiment 2, with the previously mentioned -0.20 coefficient, that confirms a correlation between higher IAT scores and lower performance ratings, though as the coefficient has been deemed statistical insignificant the hypothesis cannot be validated by this experiment alone.

The other two main effects shown are the information and advertisement variables, which have the coefficients of 1.20 and 0.51. This means that compared to the control group, the test persons which were treated with the dichotomous variable *information,* meaning that they received positive information about the USPS, show on average a 1.20 unit increase in USPS performance rating. Similar, but a bit less strong effect is seen when looking at the dichotomous variable *advertisement,* where test persons being subjected to a positive 1 minute advertisement show on average a 0.51 unit increase in the USPS performance rating. Both variables show highly statistical significance with p values of <0.000 and 0.011. Combined with their highly statistical significance they also show low SE values of 0.22 and 0.20, suggesting that our sample mean is a relatively strong representation of the true population.

The two interaction effects, *information x IAT score* and *advertisement x IAT score* both have negative coefficients of -0.13 and -0.70. Both results are detrimental to the second hypothesis H2, which argues that positive information about the USPS will attenuate, but not eliminate the influenceof antipublic sector attitudes. The coefficients show that when treated with the dichotomous variableof information and advertisement, a 1 unit higher IAT test score is on average associated with a 0.13and 0.70 unit decrease of USPS performance rating. However, again it is important to note that bothinteraction coefficients have high p-values of 0.790 and 0.115, meaning that they are statistical

insignificant. Furthermore, both interaction coefficients show higher SE values, suggesting that the results are further away from the true population than the other coefficients in the second experiment. However, both SE values are still quite low and should be considered decent representation of the true population.

Question 2.1

We are working with a one sample representative sample of the danish population and with a number of observations of the high anchored sample $n = 442$

Our null hypothesis is $H_0: p = 0.48$ and our alternative hypothesis is $H_a: p \neq 0.48$

Our test statistic is a one-sample test for a proportion, and we have an alpha value $a = 0.05$

The reference distribution is calculated

$$\bar{x} = \frac{(\sum x_i)}{n} \qquad \bar{x} = \frac{232}{442} \approx 0{,}52489$$

The standard error for significance test

$$se = \sqrt{\frac{p(1-p)}{n}} \qquad se = \sqrt{\frac{0.48(1-0.48)}{442}} \approx 0{,}023764$$

The z-score for our sample estimate

$$z - score = \frac{\bar{x_n} - p}{se\ of\ \bar{x_n}} \qquad z - score = \frac{0{,}52489 - 0{,}48}{0{,}023764} \approx 1{,}889$$

Now we calculate the p-value using R $P = 0{,}05889183$

Since our p-value is greater than our alpha of 0.05 we will retain / fail to reject our null hypothesis H_0, that the proportion of women in the high anchor group is exactly 0.48. If we change our alpha value to 0.1 instead of 0.05, the null hypothesis H_0 will be rejected as $p = 0{,}0589 < a = 0{,}1$.With this alpha level we can say that there is statistical significance to say that the proportion of women in the high anchored group are on average different from 0.48.

First by conducting a no continuity correction prop.test with a 95% confidence level in R we get the $p - value = 0.05891$, the same as in the manual calculations. If we, however, do use the continuity correction prop.test with 95% confidence interval we get a slightly more conservative estimation and

$p - value = 0.06558$, which is slightly higher than our manually calculated p-value.

Question 2.2

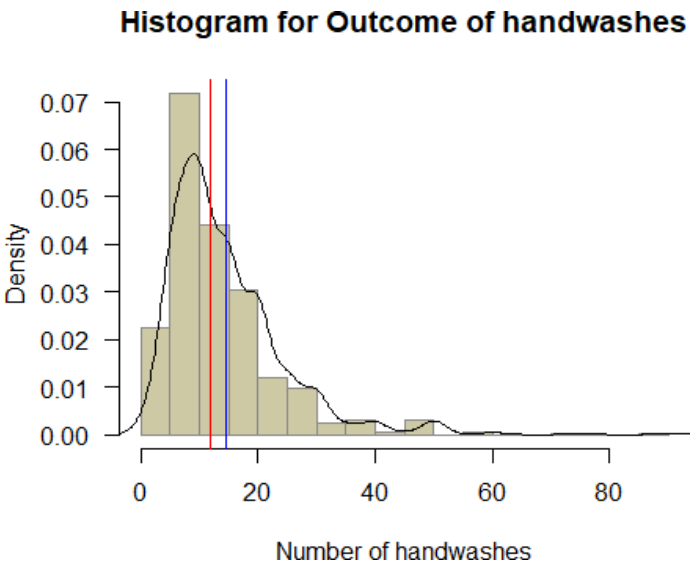
First, we calculate the average age of government supporters in the low anchor group $la.sup.age = 54.2342$ Second, we calculate the average age of opposition supporters in the low anchor group $la.opo.age = 55.47399$. We assume that all assumptions regarding interval scale and representative large sample size are met. Our null hypothesis H_0: is $\mu_0 = \mu_1$, and og $H_a: \mu_0 \neq \mu_1$ meaning that the average age of government supports in the low anchored group is statistically not different than the average age of government oppositions in the low anchored group. We have chosen an alpha level of 0.05 and perform a two sample t.test in R, which gives us that $p = 0{,}4895$

Because $p > a$ we fail to reject H_0, meaning that the average age of the government supporters in the low anchored group is not statistical significantly different from the average age of the government oppositions in the low anchored group. As the p-value already is greater than our alpha, lowering alpha to 0,01 does not chance the conclusion that we fail to reject the H_0

Question 2.3

Looking into the variable "outcome_handwash" we find some important values describing its central tendency and spread. The starting value of the central tendency is the actual midpoint of the observations, which for this variable is $median = 12$, however this value can seem arbitrary without knowledge of the range of values it is a part of. Therefore, we see that the variable ranges from $\min = 0$ to $\max = 90$, suggesting that the variable data might be skewed to the right. This is also supported by our second important central tendency measure, the mean, which in this data is $mean = 14.64$. The fact that the mean is higher than the median, further suggest a rightwards skewing of our variable's data. To get a better overview of the spread of our data we can use the Inter quartile range (IQR) function to figure out what the values are of our most central data, from the 25th percentile to the 75th percentile. In this case the IQR is 12, with a 1st quantile of 8 and a 3rd quantile of 20. This means that 50% of our data is within the 12 unit span from 8 to 20. Backed up by a 10th quantile function, showing that 80% of the data lies within 20 number of handwashes, we observe a variable which are highly skewed to the right and have some outliers past the 25 unit mark, extending the range to 90. Furthermore, the Standard deviation (SD) can elaborate on the spread of the data. The $\sigma = 9.906873$, again supporting our findings of a densely packed dataset from 0 - 25 handwashes. Quickly looking at the $2\sigma = 19.81375$, we see that there must be very few observations between 35 units and 90 units of handwashing.

To the right a univariate summary plot, a histogram of the variable is seen. By quickly looking at we can clearly recognize the rightwards skewing that our previous data suggested. The mean (blue) is higher than our median (red) and we can clearly see that very few observations lie after 30, meaning that we have a densely packed graph from 0 - 30 and that the range from 30 - 90 mostly consists of non-numerous outliers.



To calculate the standard error of the mean (SE) we use the following formula, as we are looking at a sample mean, we use the sample standard deviation, which was calculated earlier.

$$SE = \frac{s}{\sqrt{n}} \qquad SE = \frac{9.906873}{\sqrt{884}} \approx 0{,}3332$$

We, now how our standard error of the mean, and to increase its usefulness we calculate the confidence intervals (CI).

$$CI_{1-a} = [\hat{\beta} - z^*_{a/2} \times standard\ error, \hat{\beta} + z^*_{a/2} \times standard\ error]$$

We construct a 95% confidence interval, which gives us

$$z^*_{a/2} = 1{,}96$$

$$CI_{0.95} = [14{,}64253 - 1{,}96 \cdot 0{,}3332, 14{,}64253 + 1{,}96 \cdot 0{,}3332]$$

$$CI_{0{,}95} = [13.989, 15.296]$$

The standard error of 0,3332 is relatively low, meaning that our data sample mean can be considered a decent representation of the true population mean. Having calculated our CI we can say that with a 95% confidence interval that if the sampling is done enough times, 95% of the times the true value of the mean will be within the interval of $[13.989, 15.296]$.

Question 2.4

Calculating the different group means, we observe that the overall mean $= 14.64253$, the low anchored mean $= 10.85747$ and the high anchored mean $= 18.4276$. Considering the questions used for the dichotomous anchored variable, which had extremes of 3 or 30, none of the means in their respective anchoring, are close to that. However, it is clear that the dichotomous anchoring variable did have an effect, as the means in both the anchoring treatments lay roughly 4 units above or below the overall mean, clearly indicating that the anchor treatment did succeed in sorting the self-reported hand washing count. Furthermore, a regression model is fitted in R which gives us that $intercept = 18.4276$ and that $treat_{handwashlow} = -7.570136$. As previously mentioned, our anchoring treatment is a dichotomous variable, meaning that our intercept is representing a mean for the non-low anchor treatment (high), which is the same as previously seen. The beta coefficient low anchor treatment suggests that when treated with this variable, the mean is reduced by 7,57, which gives us the previously mentioned mean from the low anchored group, 10,86.

We can formulate the speculation of the anchoring treatment effect as a hypothesis test with a null hypothesis, stating that the variable had no effect. Utilizing the summary function on our regression model in R, we find that the p-value of the low anchor handwash treatment is 2e-16, which is very very small, and enables us to reject the null hypothesis, suggesting that the treatment did have an effect, and that the variable is very statistically significant. Furthermore, this is supported by a 95% CI of [-8.779409, -6.360862], where the no effect (slope value of 0) does not appear, double checking with a 99% CI, we get [-9.160652 -5.97962] which further supports our argument. We observe a SE of our low anchor coefficient of $SE = 0{,}6161$ which indicates that our sample variable mean is a good indicator of the true population mean. Evidently, we can see with great certainty that the handwashing anchor treatment did make a difference.

Question 2.5

The anchoring treatment's effect on the number of close contacts is investigated in R, in the same way as the effect on the self-reported number of handwashes. The results are: Overall mean$= 7.566$, the high anchored mean $= 8.423$ and low anchored mean $= 6.708$. Again, suggesting that the anchoring treatment did make a difference. The regression model fitted is $intercept = 8.423$, $treat_{handwashlow} = -1.715$. Which means that the groups that received the high anchor treatment had a mean of 8.423 and that the mean decreases by 1.715 for the low anchored groups, resulting in a mean of 6.708. Continuing with a null hypothesis of no effect, we see a p-value of 0,0215, meaning

that with a 95% CI we reject the null hypothesis, suggesting a clear effect off the treatment. This is supported by our $CI_{0.95} = [-3.176170, -0.2536944]$ which does not include 0, a no effect of the coefficient. The standard error is $SE = 0.7445$, meaning that our treatment sample average is a fairly good representation of the true population mean.

My expectations would correctly be that the anchoring treatment did have an effect on the number of close contacts. My rationalization would be that a high anchored handwashing variable would capture the people most worried of COVID19 and therefore also represent the people who would take stronger social distancing measures, reducing the numbers of close contacts. However, the data presented actually show the complete opposite. The treatment did have an effect, but it is suggestingthat the people anchored with the low level of handwashing on average has 1.7 fewer close contacts,struggling to believe the rationalization of people reducing their close contacts because they have notwashed their hands "enough", I do not have a ready explanation for the trend.

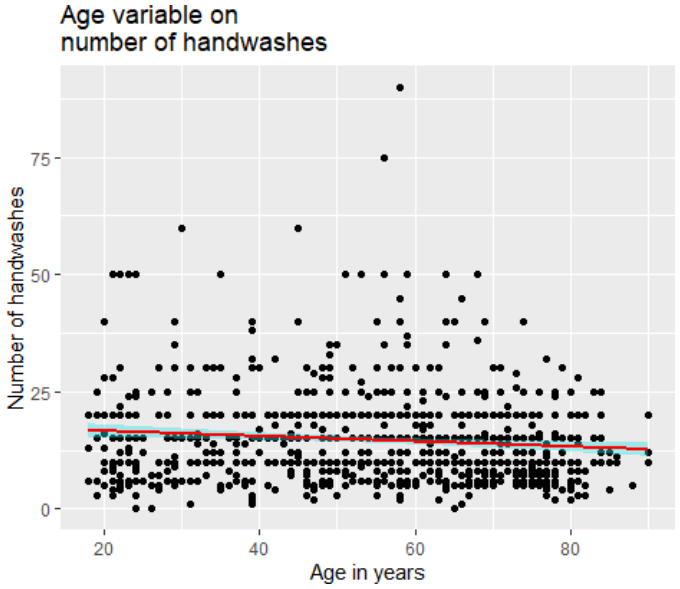
Question 2.6

The model has been extended in R with the predictors and the coefficients are shown in the table.

	Intercept	*Treat_handwaslow*	*Age*	*Male*	*Gov*
Coefficients	22.52728	-7.61792	-0.05431	-2.64371	0.25971
p-value	<2e-16	<2e-16	0.00115	1.59e-05	0.675
Std. Error	1.1251	0.607	0.01665	0.60914	0.61917
$CI_{0.95}$	[20.32, 24.74]	[-8.81, - 6.43]	[-0.09, -0.02]	[-3.84, - 1.45]	[-0.96, 1.47]

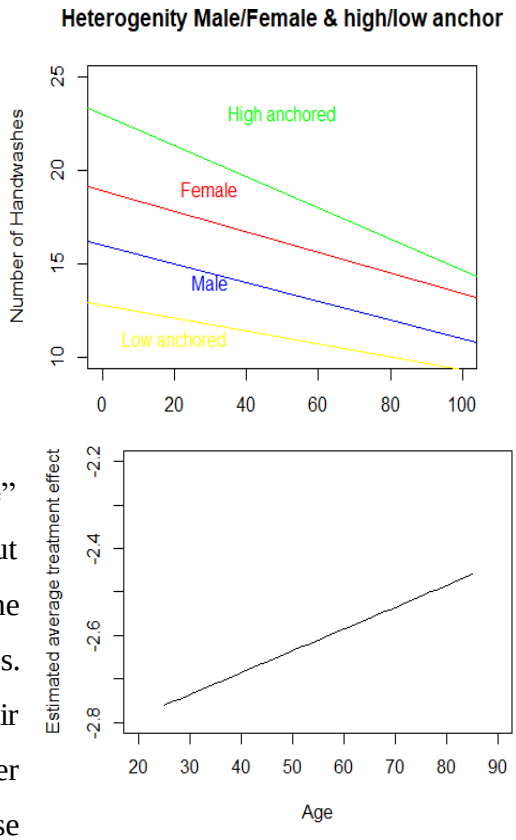
The intercept tells us that this model predicts that the average test person, who is treated with the high anchored handwash variable, theoretical 0 years of age, is a female and supports the opposition has an average of 22.52728 handwashes the day before they answered the questionnaire. If instead a person belongs in the low anchored handwash variable their number of handwashes on average decrease with the "Treat_handwashlow" coefficient of 7.62 handwashes pr. Day. The age variable estimates that on average the test person will wash their hand 0.0543 times less pr. day for every year older they become. The "male" coefficient of -2.644 tells us that if a test subject is male their average number of handwashes pr. day decreases by 2.644. The "gov" coefficient describes how the number of handwashes on average increases by 0.26 if the test subjects are supporters of the government. With the exception of the government variable, all variables show a very low p-value indicating

strong statistical significance and also a relatively low standard error, suggesting the results are representative of the true population mean. This is generally supported by our presented Ci's which show that the coefficients have a relatively certain effect, with the exception of the government variable. Other than having a small p-value, its CI ranges from negative to positive, meaning that 95% of the times thetrue value could very well be both positive and negative, rendering our gov coefficient very uncertain for conclusions. To the right the effect of age is shownwith a red decreasing regression coefficient and a lightblue colored area around the regression to show the CI



Question 2.7

A heterogeneous treatment effect is a treatment effect that only has an effect when a particular condition is met. In this case I have chosen to model the Male/Female and the Low/high anchor treatment, utilizing their specific interaction terms, in a basic comparison graph, but also a specific estimated treatment effect graph for the Male/Female treatment. This is done to easily identify any heterogeneity for different ages. In the top graph to the right the blue and red line illustrates the effect of being treated with the "male" variable on different age groups. The male line is shifted downwards by 2.9 units suggesting the base effect of the treatment. However, the male and female lines run substantially parallel and linear suggesting, that the treatment effect is almost identical for all different ages. The bottom graph zooms further in on this, by showing the estimated "Male" treatment effect for the different ages. An effect of age is seen but relatively very small. Furthermore, the green and the yellow line illustrates the High/Low anchoring treatment effect on different ages. The graphs are also perfectly linear with a base treatment effect of their difference of intercept; however, we can observe a tendency of higher treatment effect the younger the test subject was, and a stable decrease in treatment effect the older they become.



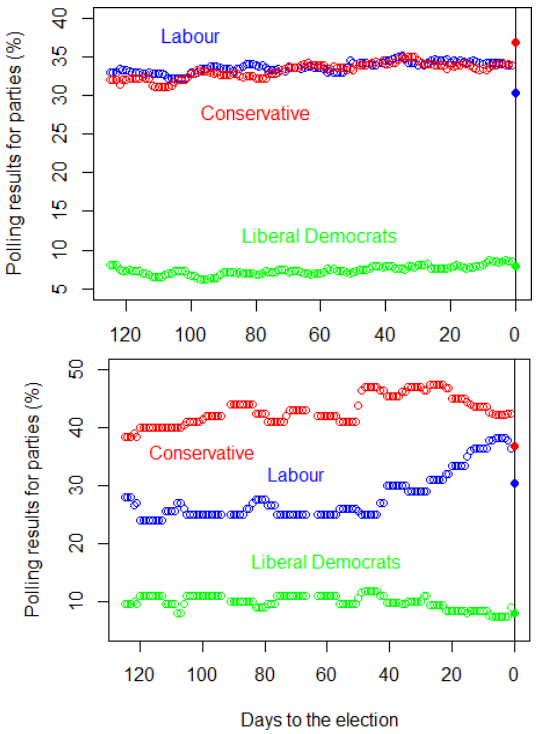
Question 3.1

As seen in the table to the right the Opinium had the lowest average prediction error (Bias) in 2/3 parties in 2015 and in alle three parties in 2017. They also had the smallest error in their prediction closest to the election in all 3 parties in 2015 and in 2/3 parties in 2017. Furthermore, the average BIAS for Opinium was 2,823 and the average RMSE was 3,758, compared to YouGov's average BIAS of 4,141 and their average RMSE of 4,423. Evidently, Opinium did better than YouGov. There is no doubt that the labour party was the most difficult party to predict, especially in 2017, where both polling companies we extremely wrong with the average predictions. This is further illustrated with the average RMSE for The Labour party being 7,17, the Conservatives 3,48 and the Liberal Democrats 1,63. In 2015 the Labour and the conservative party were roughly equally hard to predict for the two companies. Ignoring YouGov's 2017 average poll, both companies were quite successful in predicting the Liberal democrats votes in both years.

Lab / Con / Libdem	*Mean predict*	*Latest predict*	*Results*	*Bias*	*Latest PE*	*RMSE*
Opinium 2015	33.59	34	30.4	3.19	3.6	3.3
YouGov 2015	33.81	34	30.4	3.4	3.6	3.61
Opinium 2017	31	36	40.0			9.58
YouGov 2017	28.9	35	40.0	-11.1	-5	12.19
Opinium 2015	33.65	35	36.9	-3.25	-1.9	3.79
YouGov 2015	33.32	34	36.9	-3.58	-2.9	3.79
Opinium 2017	42.77	43	42.4	0.37	0.6	3.49
YouGov 2017	43.19	42	42.4	-3.58	-0.4	2.84
Opinium 2015	7.29	8	7.9	0.61	0.1	1.13
YouGov 2015	7.5	9	7.9	-0.4	1.1	1.04
Opinium 2017	7.92	8	7.4	0.52	0.6	1.26
YouGov 2017	10.19	10	7.4	2.79	2.6	3.07

Question 3.2

The top graph to the right shows the 2015 election, where change in polling results as the election comes closer. As we near two months before the election, the changes for the Labour and conservative party are quite small, and their predictions do not get any nearer the actual result. The predictions for the Labour party has RMSE values for the 125 to 80 days to election period of 2.98, days 80 to 40 of 3.6 and days 40 to 0 of 3.99, suggesting that the predictions got worse the closer to election. The RMSE values for the same interval periods for conservative was 4.79, 3.34 and 3.04. Suggesting the opposite trend, increasing quality closer to the election. For the Liberal democrats, the pattern is almost the same, however, a few days before the election they predict a quite substantial decrease, which can be seen was correct. The RMSE



interval values was 1.29, 1.00 and 0.83, showing a pattern that resembles the conservative with increasing quality of predictions closer to the election. Comparing to the bottom graph representing the 2017 election we observe much more dramatic changes for especially Labour at the 40 days mark, however this dramatic increase was a wrongful prediction, as the actual outcome were more similar to that of the 30 day mark. The RMSE values for the interval periods were 14.44, 13.7 and 6.87, though very high RMSE values, a pattern of increased quality closer to election does appear. The Conservative party prediction does not chance much through the first 70 days, however a substantial rise in seen from 50 days before election until 30 days before, where the poll companies decrease the prediction, however, not enough. The RMSE values for the interval periods were 1.82, 3.36 and 3.11, suggesting a somewhat decrease of quality as we got closer to the election Again, the closest to election prediction quality of the Liberal democrats is much better than of the two other parties. They have RMSE values for the interval periods of 2.87, 3.51 and 1.81, indicating an increasing quality of prediction the fewer days there were to the election day,

Question 3.3

Polls with a large sample size are defined as N>2000 and small sample sizes defined as N<1750. The Bias scores and RMSE scores are shown for both years, and for all parties. Both sample sizes in the table below. In both time periods the larger samples sizes seem to be more accurate as they have the lowest bias score and RMSE scores in 5/6 possible predictions. This could be explained by the law of large numbers which states that when a sample size increases its mean converges to the true mean of the population. As the bias scores are a measure of the error of the true mean (election results), a consistent lower bias, means a closer prediction to the population mean, which the LLN argues should come from the largest sample sizes. This is the case with our data, in 5/6 predictions.

Bias/RMSE	*Large N (Lab)*	*Small N(Lab)*	*Large N (Con)*	*Small N(Con)*	*Large N (Libdem)*	*Small N(Libdem)*
2015	3.78	3.26	-2.9	-3.74	0.19	-0.64
2017	-9.35	-13.29	0.45	1.13	1.35	3.31
2015	3.83	3.49	3.1	3.97	0.69	1.17
2017	10.32	13.49	3.17	3.36	2.08	3.39

Bibliography

Lange, R. T. (2011). Inter-rater Reliability. In *Encyclopedia of Clinical Neuropsychology* (pp. 1348–1348). Springer New York. https://doi.org/10.1007/978-0-387-79948-3_1203

Staller, K. M. (2010). Encyclopedia of Research Design - Experimental Design. *Encyclopedia of Research Design: Qualitative Research*, 448–453.

R - coding Appendix

```
#  Final exam 01/06 - 2021

#       Question 2.1

covid <- read.csv("covid.csv")

dim(covid)

names(covid)


covid.ha <- subset(covid, treat_handwash == "high")

nrow(covid.ha)

sum(covid.ha$male == 0)

xbar <- 0.52489

se <- 0.023764

zscore <- 1.889

2*pnorm(zscore, lower.tail = FALSE)

prop.test(232, n = 442, p = 0.48, conf.level = 0.95, correct = FALSE)

prop.test(232, n = 442, p = 0.48, conf.level = 0.95



#       Question 2.2
```

```
covid.la <- subset(covid, treat_handwash == "low")
la.sup <- subset(covid.la, gov == 1)
la.opo <- subset(covid.la, gov == 0)
la.sup.age <- mean(la.sup$age)
la.opo.age <- mean(la.opo$age)
t.test(la.opo$age, la.sup$age)

#       Question 2.3
summary(covid$outcome_handwash)
IQR(covid$outcome_handwash)
quantile(covid$outcome_handwash, probs = seq(from = 0, to = 1, by = 0.1))
sd(covid$outcome_handwash)
2*sd(covid$outcome_handwash)
hist(covid$outcome_handwash,
    breaks = 20,
    main = "Histogram for Outcome of handwashes",
    xlab = "Number of handwashes",
    border = "snow4",
    col = "lemonchiffon3",
    las = 1,
    prob = TRUE
)
lines(density(covid$outcome_handwash))
```

```
abline(v = mean(covid$outcome_handwash), col = "blue")
abline(v = median(covid$outcome_handwash), col = "red")

#       Question 2.4

mean.covid.ha <- mean(covid.ha$outcome_handwash)
mean.covid.la <- mean(covid.la$outcome_handwash)
mean.covid <- mean(covid$outcome_handwash)
m1 <- lm(outcome_handwash ~ treat_handwash, data = covid)
coef(m1)
summary(m1)
confint(m1)
confint(m1, level = 0.99)

#       Question 2.5
mean.cc.covid.ha <- mean(covid.ha$outcome_closecontact)
mean.cc.covid.la <- mean(covid.la$outcome_closecontact)
mean.cc.covid <- mean(covid$outcome_closecontact)
m2 <- lm(outcome_closecontact ~ treat_handwash, data = covid)
coef(m2)
summary(m2)
confint(m2)
```

Question 2.6

```
m3 <- lm(outcome_handwash ~ treat_handwash + age + male + gov, data = covid)
coef(m3)
summary(m3)
mean(covid$outcome_handwash)

mean(covid$outcome_handwash[covid$male == 1])
mean(covid$outcome_handwash[covid$male == 0])

mean(covid$outcome_handwash[covid$age < 35])
mean(covid$outcome_handwash[covid$age > 35])

mean(covid$outcome_handwash[covid$gov == 1])
mean(covid$outcome_handwash[covid$gov == 0])

install.packages("texreg")
library("texreg")
screenreg(m3)

confint(m3)

install.packages("ggplot2")
library("ggplot2")
```

```
install.packages("hrbrthemes")
library(hrbrthemes)

age.plot <- ggplot(covid, aes(x = age, y = outcome_handwash)) +
      geom_point()+
      geom_smooth(method = lm, color = "red", fill = "69b3a2", se = TRUE)
theme_ipsum()

print(age.plot + ggtitle("Age variable on \nnumber of handwashes") +
            labs(y= "Number of handwashes", x = "Age in years"))

#       Question 2.7

m5 <- lm(outcome_handwash ~ age * male, data = covid)
coef(m5)

m7 <- lm(outcome_handwash ~ age * treat_handwash, data = covid)
coef(m7)

age.male <- data.frame(age = seq(from = 25, to = 85, by = 20),
                       male = 1)

age.female <- data.frame(age = seq(from = 25, to = 85, by = 20),
                       male = 0)
```

```
ate.age <- predict(m5, newdata = age.male) -
      predict(m5, newdata = age.female)

yt.hat <- predict(m5,
            newdata = data.frame(age = 25:85, male = 1))

yc.hat <- predict(m5,
            newdata = data.frame(age = 25:85, male = 0))

plot(x = 25:85, y = yt.hat - yc.hat, type = "l", xlim = c(20, 90),
    ylim = c(-2.8, -2.2),
    xlab = "Age",
    ylab = "Estimated average treatment effect")

plot(1, type = "n",
    xlab = "Age in Years", ylab = "Number of Handwashes",
    main = "Heterogenity Male/Female & high/low anchor",
    xlim = c(0, 100), ylim = c(10, 25))

abline(a = 18.89797273, b = -0.05512306, col = "red")
abline(a = (18.89797273 - 2.88700988), b = (-0.05512306 + 0.00502327), col = "blue" )
text(30, 14, "Male", col = "blue")
text(30, 19, "Female", col = "red")
```

```
abline(a = 23.01966797, b = -0.08339606, col = "green")
abline(a = (23.01966797 - 10.27060570), b = (-0.08339606  +  0.04882707), col = "yellow")
text(20, 11, "Low anchored", col = "yellow")
text(50, 23, "High anchored", col = "green")


#       Question 3.1

load("poll-uk.Rdata")
results
dim(polls)

install.packages("Metrics")
library("Metrics")

rmse(poll.Opinium.2015$vote_lab, results$res_lab[results$election == "2015"])
rmse(poll.Opinium.2015$vote_con, results$res_con[results$election == "2015"])
rmse(poll.Opinium.2015$vote_libdem, results$res_libdem[results$election == "2015"])
rmse(poll.Yougov.2015$vote_lab, results$res_lab[results$election == "2015"])
rmse(poll.Yougov.2015$vote_con, results$res_con[results$election == "2015"])
rmse(poll.Yougov.2015$vote_libdem, results$res_libdem[results$election == "2015"])

rmse(poll.Opinium.2017$vote_lab, results$res_lab[results$election == "2017"])
```

```
rmse(poll.Opinium.2017$vote_con, results$res_con[results$election == "2017"])

rmse(poll.Opinium.2017$vote_libdem, results$res_libdem[results$election == "2017"])

rmse(poll.Yougov.2017$vote_lab, results$res_lab[results$election == "2017"])

rmse(poll.Yougov.2017$vote_con, results$res_con[results$election == "2017"])

rmse(poll.Yougov.2017$vote_libdem, results$res_libdem[results$election == "2017"])



unique(polls$house)

polls$newdate <- as.Date(polls$poll_date)


poll.Opinium <- subset(polls, house == "Opinium")

poll.Opinium.2015 <- subset(poll.Opinium, poll_date >= "2015-01-02" & poll_date <= "2015-05-05")

poll.Opinium.2017 <- subset(poll.Opinium, poll_date > "2015-05-05")


pred.2015.Opinium.lab <- mean(poll.Opinium.2015$vote_lab)

pred.2015.Opinium.con <- mean(poll.Opinium.2015$vote_con)

pred.2015.Opinium.libdem <- mean(poll.Opinium.2015$vote_libdem)


pred.2015.Opinium.lab.latest <- poll.Opinium.2015$vote_lab[poll.Opinium.2015$poll_date == "2015-05-05"]

pred.2015.Opinium.con.latest <- poll.Opinium.2015$vote_con[poll.Opinium.2015$poll_date == "2015-05-05"]

pred.2015.Opinium.libdem.latest <-poll.Opinium.2015$vote_libdem[poll.Opinium.2015$poll_date == "2015-05-05"]
```

```
mean.error.lab <- pred.2015.Opinium.lab - results$res_lab[results$election == "2015"]

mean.error.con <- pred.2015.Opinium.con - results$res_con[results$election == "2015"]

mean.error.libdem <- pred.2015.Opinium.libdem - results$res_libdem[results$election == "2015"]


pred.2015.Opinium.lab.latest - results$res_lab[results$election == "2015"]

pred.2015.Opinium.con.latest - results$res_con[results$election == "2015"]

pred.2015.Opinium.libdem.latest- results$res_libdem[results$election == "2015"]


pred.2017.Opinium.lab <- mean(poll.Opinium.2017$vote_lab)

pred.2017.Opinium.con <- mean(poll.Opinium.2017$vote_con)

pred.2017.Opinium.libdem <- mean(poll.Opinium.2017$vote_libdem)


pred.2017.Opinium.lab.latest <- poll.Opinium.2017$vote_lab[poll.Opinium.2017$poll_date ==
"2017-06-06"]

pred.2017.Opinium.con.latest <- poll.Opinium.2017$vote_con[poll.Opinium.2017$poll_date ==
"2017-06-06"]

pred.2017.Opinium.libdem.latest <-poll.Opinium.2017$vote_libdem[poll.Opinium.2017$poll_date
== "2017-06-06"]


mean.error.lab <- pred.2017.Opinium.lab - results$res_lab[results$election == "2017"]

mean.error.con <- pred.2017.Opinium.con - results$res_con[results$election == "2017"]

mean.error.libdem <- pred.2017.Opinium.libdem - results$res_libdem[results$election == "2017"]


pred.2017.Opinium.lab.latest - results$res_lab[results$election == "2017"]
```

```
pred.2017.Opinium.con.latest - results$res_con[results$election == "2017"]

pred.2017.Opinium.libdem.latest- results$res_libdem[results$election == "2017"]


poll.Yougov <- subset(polls, house == "YouGov")

poll.Yougov.2015 <- subset(poll.Yougov, poll_date >= "2015-01-05" & poll_date <= "2015-05-05")

poll.Yougov.2017 <- subset(poll.Yougov, poll_date > "2015-05-05")


pred.2015.Yougov.lab <- mean(poll.Yougov.2015$vote_lab)

pred.2015.Yougov.con <- mean(poll.Yougov.2015$vote_con)

pred.2015.Yougov.libdem <- mean(poll.Yougov.2015$vote_libdem)


pred.2015.Yougov.lab.latest <- poll.Yougov.2015$vote_lab[poll.Yougov.2015$poll_date == "2015-05-05"]

pred.2015.Yougov.con.latest <- poll.Yougov.2015$vote_con[poll.Yougov.2015$poll_date == "2015-05-05"]

pred.2015.Yougov.libdem.latest <-poll.Yougov.2015$vote_libdem[poll.Yougov.2015$poll_date == "2015-05-05"]


mean.error.lab.Yougov <- pred.2015.Yougov.lab - results$res_lab[results$election == "2015"]

mean.error.con.Yougov <- pred.2015.Yougov.con - results$res_con[results$election == "2015"]

mean.error.libdem.Yougov <- pred.2015.Yougov.libdem - results$res_libdem[results$election == "2015"]


pred.2015.Yougov.lab.latest - results$res_lab[results$election == "2015"]
```

```
pred.2015.Yougov.con.latest - results$res_con[results$election == "2015"]

pred.2015.Yougov.libdem.latest- results$res_libdem[results$election == "2015"]


pred.2017.Yougov.lab <- mean(poll.Yougov.2017$vote_lab)

pred.2017.Yougov.con <- mean(poll.Yougov.2017$vote_con)

pred.2017.Yougov.libdem <- mean(poll.Yougov.2017$vote_libdem)


pred.2017.Yougov.lab.latest <- poll.Yougov.2017$vote_lab[poll.Yougov.2017$poll_date == "2017-
06-07"]

pred.2017.Yougov.con.latest <- poll.Yougov.2017$vote_con[poll.Yougov.2017$poll_date ==
"2017-06-07"]

pred.2017.Yougov.libdem.latest <-poll.Yougov.2017$vote_libdem[poll.Yougov.2017$poll_date ==
"2017-06-07"]


mean.error.lab.Yougov <- pred.2017.Yougov.lab - results$res_lab[results$election == "2017"]

mean.error.con.Yougov <- pred.2017.Yougov.con - results$res_con[results$election == "2017"]

mean.error.libdem.Yougov <- pred.2017.Yougov.libdem - results$res_libdem[results$election ==
"2017"]


pred.2017.Yougov.lab.latest - results$res_lab[results$election == "2017"]

pred.2017.Yougov.con.latest - results$res_con[results$election == "2017"]

pred.2017.Yougov.libdem.latest- results$res_libdem[results$election == "2017"]


#       Question 3.2
```

```
poll.2015 <- subset(polls, poll_date >= "2015-01-01" & poll_date <= "2015-12-30")
poll.2015$DaystoElection <- as.Date("2015-05-07") - poll.2015$poll_date

vote_lab.pred <- vote_con.pred <- vote_libdem.pred <- rep(NA, 125)
for (i in 1:125) {
      week.data <- subset(poll.2015, subset = ((DaystoElection <= (125 - i + 7))
                  & (DaystoElection > (125 - i ))))
      vote_lab.pred[i] <- mean(week.data$vote_lab)
      vote_con.pred[i] <- mean(week.data$vote_con)
      vote_libdem.pred[i] <- mean(week.data$vote_libdem)
}

plot(125:1, vote_lab.pred, type = "b", xlim = c(125, 0), ylim = c(5, 40),
    col = "blue", xlab = "Days to the election",
    ylab = "Polling results for parties (%)")
lines(125:1, vote_con.pred, type = "b", col = "red")
lines(125:1, vote_libdem.pred, type = "b", col = "green")
abline(v = 0)
points(0,30.4, pch = 19, col = "blue")
points(0,36.9, pch = 19, col = "red")
points(0,7.9, pch = 19, col = "green")
text(100, 38, "Labour", col = "blue")
text(80, 28, "Conservative", col = "red")
text(60, 12, "Liberal Democrats", col = "green")
```

```
poll.2017 <- subset(polls, poll_date > "2017-01-01")
poll.2017$DaystoElection <- as.Date("2017-06-08") - poll.2017$poll_date

vote_lab.pred <- vote_con.pred <- vote_libdem.pred <- rep(NA, 125)
for (i in 1:125) {
      week.data <- subset(poll.2017, subset = ((DaystoElection <= (125 - i + 7))
                                    & (DaystoElection > (125 - i ))))
      vote_lab.pred[i] <- mean(week.data$vote_lab)
      vote_con.pred[i] <- mean(week.data$vote_con)
      vote_libdem.pred[i] <- mean(week.data$vote_libdem)
}

plot(125:1, vote_lab.pred, type = "b", xlim = c(125, 0), ylim = c(5, 50),
    col = "blue", xlab = "Days to the election",
    ylab = "Polling results for parties (%)")
lines(125:1, vote_con.pred, type = "b", col = "red")
lines(125:1, vote_libdem.pred, type = "b", col = "green")
abline(v = 0)
points(0,30.4, pch = 19, col = "blue")
points(0,36.9, pch = 19, col = "red")
points(0,7.9, pch = 19, col = "green")
text(100, 36, "Conservative", col = "red")
text(70, 32, "Labour", col = "blue")
```

```
text(60, 17, "Liberal Democrats", col = "green")


rmse(poll.2015$vote_lab[poll.2015$DaystoElection >= 80 & poll.2015$DaystoElection <= 125],
    results$res_lab[results$election == "2015"])
rmse(poll.2015$vote_lab[poll.2015$DaystoElection >= 40 & poll.2015$DaystoElection < 80],
    results$res_lab[results$election == "2015"])
rmse(poll.2015$vote_lab[poll.2015$DaystoElection >= 0 & poll.2015$DaystoElection < 40],
    results$res_lab[results$election == "2015"])
rmse(poll.2015$vote_con[poll.2015$DaystoElection >= 80 & poll.2015$DaystoElection <= 125],
    results$res_con[results$election == "2015"])
rmse(poll.2015$vote_con[poll.2015$DaystoElection >= 40 & poll.2015$DaystoElection < 80],
    results$res_con[results$election == "2015"])
rmse(poll.2015$vote_con[poll.2015$DaystoElection >= 0 & poll.2015$DaystoElection < 40],
    results$res_con[results$election == "2015"])
rmse(poll.2015$vote_libdem[poll.2015$DaystoElection >= 80 & poll.2015$DaystoElection <=
125],
    results$res_libdem[results$election == "2015"])
rmse(poll.2015$vote_libdem[poll.2015$DaystoElection >= 40 & poll.2015$DaystoElection < 80],
    results$res_libdem[results$election == "2015"])
rmse(poll.2015$vote_libdem[poll.2015$DaystoElection >= 0 & poll.2015$DaystoElection < 40],
    results$res_libdem[results$election == "2015"])


rmse(poll.2017$vote_lab[poll.2017$DaystoElection >= 80 & poll.2017$DaystoElection <= 125],
    results$res_lab[results$election == "2017"])
```

```
rmse(poll.2017$vote_lab[poll.2017$DaystoElection >= 40 & poll.2017$DaystoElection < 80],
    results$res_lab[results$election == "2017"])
rmse(poll.2017$vote_lab[poll.2017$DaystoElection >= 0 & poll.2017$DaystoElection < 40],
    results$res_lab[results$election == "2017"])
rmse(poll.2017$vote_con[poll.2017$DaystoElection >= 80 & poll.2017$DaystoElection <= 125],
    results$res_con[results$election == "2017"])
rmse(poll.2017$vote_con[poll.2017$DaystoElection >= 40 & poll.2017$DaystoElection < 80],
    results$res_con[results$election == "2017"])
rmse(poll.2017$vote_con[poll.2017$DaystoElection >= 0 & poll.2017$DaystoElection < 40],
    results$res_con[results$election == "2017"])
rmse(poll.2017$vote_libdem[poll.2017$DaystoElection >= 80 & poll.2017$DaystoElection <=
125],
    results$res_libdem[results$election == "2017"])
rmse(poll.2017$vote_libdem[poll.2017$DaystoElection >= 40 & poll.2017$DaystoElection < 80],
    results$res_libdem[results$election == "2017"])
rmse(poll.2017$vote_libdem[poll.2017$DaystoElection >= 0 & poll.2017$DaystoElection < 40],
    results$res_libdem[results$election == "2017"])


#       Question 3.3


range(poll.2015$sample)
mean.error.large.2015.lab <-mean(poll.2015$vote_lab[poll.2015$sample > 2000]) -
(results$res_lab[results$election == "2015"])
```

```
mean.error.small.2015.lab <-mean(poll.2015$vote_lab[poll.2015$sample < 1750]) -
(results$res_lab[results$election == "2015"])

mean.error.large.2015.con <-mean(poll.2015$vote_con[poll.2015$sample > 2000]) -
(results$res_con[results$election == "2015"])

mean.error.small.2015.con <-mean(poll.2015$vote_con[poll.2015$sample < 1750]) -
(results$res_con[results$election == "2015"])

mean.error.large.2015.libdem <-mean(poll.2015$vote_libdem[poll.2015$sample > 2000]) -
(results$res_libdem[results$election == "2015"])

mean.error.small.2015.libdem <-mean(poll.2015$vote_libdem[poll.2015$sample < 1750]) -
(results$res_libdem[results$election == "2015"])

rmse(poll.2015$vote_lab[poll.2015$sample > 2000], results$res_lab[results$election == "2015"])

rmse(poll.2015$vote_lab[poll.2015$sample < 1750], results$res_lab[results$election == "2015"])

rmse(poll.2015$vote_con[poll.2015$sample > 2000], results$res_con[results$election == "2015"])

rmse(poll.2015$vote_con[poll.2015$sample < 1750], results$res_con[results$election == "2015"])

rmse(poll.2015$vote_libdem[poll.2015$sample > 2000], results$res_libdem[results$election ==
"2015"])

rmse(poll.2015$vote_libdem[poll.2015$sample < 1750], results$res_libdem[results$election ==
"2015"])


range(poll.2017$sample)

mean.error.large.2017.lab <-mean(poll.2017$vote_lab[poll.2017$sample > 2000]) -
(results$res_lab[results$election == "2017"])

mean.error.small.2017.lab <-mean(poll.2017$vote_lab[poll.2017$sample < 1750]) -
(results$res_lab[results$election == "2017"])

mean.error.large.2017.con <-mean(poll.2017$vote_con[poll.2017$sample > 2000]) -
(results$res_con[results$election == "2017"])
```

```
mean.error.small.2017.con <-mean(poll.2017$vote_con[poll.2017$sample < 1750]) -
(results$res_con[results$election == "2017"])

mean.error.large.2017.libdem <-mean(poll.2017$vote_libdem[poll.2017$sample > 2000]) -
(results$res_libdem[results$election == "2017"])

mean.error.small.2017.libdem <-mean(poll.2017$vote_libdem[poll.2017$sample < 1750]) -
(results$res_libdem[results$election == "2017"])

rmse(poll.2017$vote_lab[poll.2017$sample > 2000], results$res_lab[results$election == "2017"])

rmse(poll.2017$vote_lab[poll.2017$sample < 1750], results$res_lab[results$election == "2017"])

rmse(poll.2017$vote_con[poll.2017$sample > 2000], results$res_con[results$election == "2017"])

rmse(poll.2017$vote_con[poll.2017$sample < 1750], results$res_con[results$election == "2017"])

rmse(poll.2017$vote_libdem[poll.2017$sample > 2000], results$res_libdem[results$election ==
"2017"])

rmse(poll.2017$vote_libdem[poll.2017$sample < 1750], results$res_libdem[results$election ==
"2017"])
```