Data Essay for POL2077

660009633

23/11/2019

# Introduction

I am looking to demonstrate my understanding of simple linear regression as well as a knowledge of *R*, *Markdown*, *BibTex* and various statistical R packages. I am going to utilise the *2017 British Election Study* dataset that has been aggregated to the constituency level.

Where I have a small, blue bracketed number this indicates code and its respective output position in the appendix.

I will start by calling on the various packages I will be using as well as importing the data which I have downloaded from ELE.

###### (1)

## Section 1

### a

I will be looking at *Con17* (baker\_general\_2019) as my dependent variable and *Con15*, *Con10* (noauthor\_past\_nodate) and *c11EthnicityWhite* (cracknell\_census\_2014) as my independent variables. The chosen variables are all continuous, interval variables based between 0 and 100 as they represent percentages. This means there is no requirement to standardise the data as they are all based on the same measurement. *Con17*, *Con15* and *Con10* are the Conservative vote shares for the years 2017, 2015 and 2010 respectively whereas *c11EthnicityWhite* is the percentage of the constituents whose ethnicity is white.

### b

I have split my hypothesis into three parts:

: The vote share for Conservatives in 2017 can be predicted by the Conservative vote share in 2015 due to voter stability

: The vote share for Conservatives in 2017 is positively correlated with the percentage of constituents with white ethnicity

: The 2015 Conservative Party vote share is a stronger predictor on the 2017 Conservative vote share than the % of constituents with white ethnicity.

I believe to be the case as ‘The number of safe seats increased significantly in the 2017 general election. Seats won with a margin of over 50% rose from 21 seats in 2015 to 35 in 2017. Seats won with a margin between 45% and 50% also increased from 18 in 2015 to 29 in 2017.’ (tutor2u ([2019](#ref-tutor2u_safe_2019))). A party’s prototypical policies will always do well in a certain area because of the socio-economic environment and although these do change, they do not usually do so in the short term. For example, a very wealthy constituency in the UK is more likely to favour tax-breaks, typical of conservative policy, or a constituency that contains a nature reserve will likely support green policies, typical of the Green Party.

is also a valid point as historically there have been strong relationships between race and vote on a constituency level (Khan, Sveinsson, and Runnymede Trust ([2015](#ref-khan_race_2015))). As ethnicity can be considered a form of social identity I decided that it was a good variable to compare recent vote share to and I believe it will not prove to as strong an explanatory variable.

Overall however I believe due to the short, two-year gap between the last two elections as well as the rise in safe seats will be affirmed.

###### (2)

### c

There is a lot that can be potentially taken away from this model, from the temporal analysis of previous election results we can conclude with some confidence how consistent constituencies are with their election choices. If I can show strong auto-correlation with Conservative vote results then this could provide an efficient way to predict the electorate vote in the future. This could also highlight a culture of identity politics within constituencies whereby support for specific political parties is entrenched through supporter tribalism.

Identity politics, especially race, is consistently an issue in the USA, for example, the Democrats have enjoyed a huge advantage with Black voters. According to Pew Research Center, since the 1980s the lowest advantage held from this ethnicity for Democratic candidate was still above 70points (Alec ([2016](#ref-alec_behind_2016))). I believe there is a similar phenomenon occurring in British politics, one which has been exaggerated by the vote for Britain to separate from the EU in the 2016 referendum. This hugely divisive and pivotal decision that has disrupted the UK’s political environment and may have led to race playing a more significant part in the 2017 General Election.

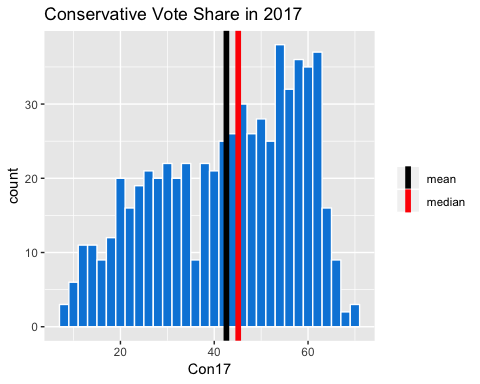
It is perhaps hyperbolic to suggest that Brexit was the genesis of the UK’s possible current social mentality however I believe it has amplified underlying ethnopolitical currents. My model will aim to quantify whether ethnicity plays a significant role in deciding the vote share for conservatives post-Brexit.

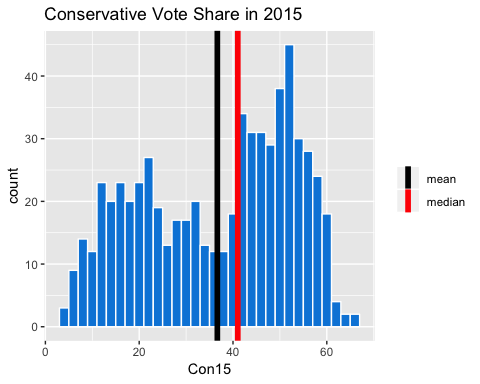
## Section 2

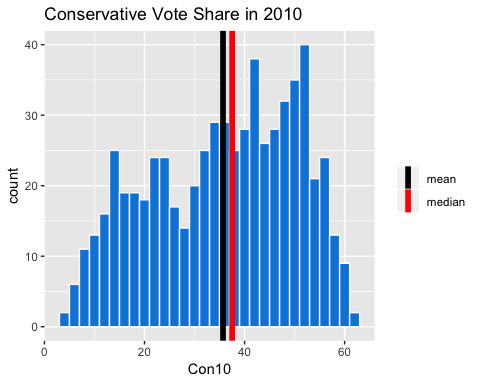
### a

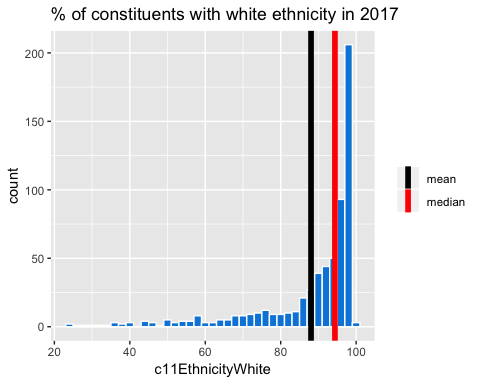
To create each of the histograms of the variables I have produced the function *histPlotR* so as not to repeat any code. The arguments are *.var* which is the variable I am plotting and *mtitle*, the main title of the plot. This function also includes mean and median vlines to show highlight the skewness of the data in these variables.

###### (3)









*Con15*, *Con10* and *c11EthnicityWhite* all show variance which is an assumption of linear regression and essential if we want these variables to be able to account for the variation in our dependent variable, *Con17*.

###### (4)

I have calculated the Kurtosis and Skewness of each graph because I believe these statistical descriptions are an important inclusion in the analysis of our variables.

The kurtosis parameter is a measure of the combined weight of the tails relative to the rest of the distribution. (Westfall ([2014](#ref-westfall_kurtosis_2014))) and it is important when assuming whether the data is normally distributed.

Where the kurtosis score is around 0 then it assumed to be normally distributed. As *Con17*, *Con15* and *Con10* are all quite negative we consider these to be ‘light-tailed’ datasets whereas *c11EthnicityWhite* is extremely ‘heavy-tailed’ which can be seen in the histogram.

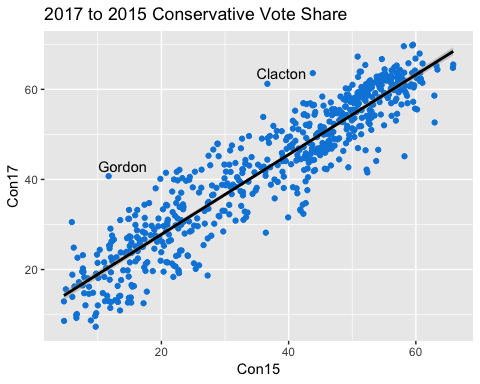
Skewness is usually described as a measure of a dataset’s symmetry – or lack of symmetry. A perfectly symmetrical data set will have a skewness of 0. The normal distribution has a skewness of 0. (McNeese ([2008](#ref-mcneese_are_2008)))

All four of my variables are negatively skewed however *Con17*, *Con15* and *Con10* again share similar scores showing them to be quite minimally skewed whereas *c11EthnicityWhite* has a heavy skew.

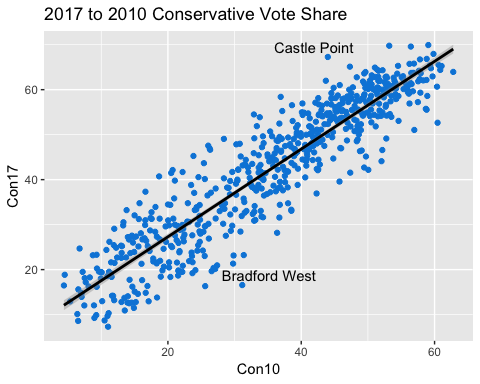
### b

To plot each of my independent variables against *Con17*, I have created a function for the same aforementioned reason of efficiency, it is also eponymously named for ease - *scatPlotR*. I have included the names of the constituencies with the largest residual values to aid potential future necessary outlier analysis.

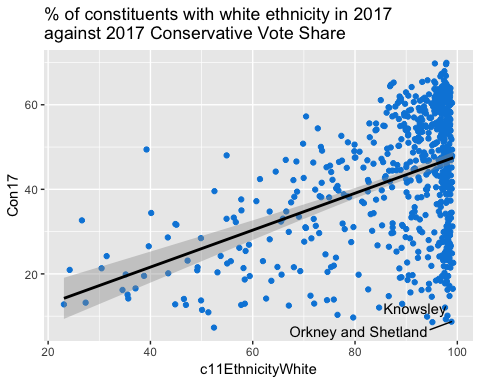
###### (5)



This graph shows a strong, linear correlation between the vote share in 2015 and 2017 for the Conservatives. This suggests that the 2015 electoral results have a positive relationship with the 2017 vote share.



This graph also shows a strong, positive relationship between the two continuous variables. There appear to be slightly higher average residuals but the correlation is still strong.



This third graph appears to have a slightly, weaker positive correlation between the variables. This could partly be to do with the unbalanced *c11EthnicityWhite* data which leads to a wider confidence interval on the left-hand side of the plot.

### c

###### (6)

Above I have calculated the correlation table between the variables rounded to two decimal places.

###### (7)

Here I have calculated the shared variance table between the variables rounded to three decimal places. For reference if the shared variance is above 0.5 then there is a problem of collinearity.

As the model I’m going to be running is multivariate, it must satisfy all ten of the assumptions in multiple regression. As this currently shows there is strong evidence of multicollinearity between *Con15* and *Con10*, this is expected as they are measurements of the same variable at two close points in time. Both are measuring conservative votes for the same leader, David Cameron, with both producing very similar manifestos.

Due to the result of the correlation and shared variance tables, it would make sense to omit one of *Con15* or *Con10*. If I do not action this then I should expect to see higher standard errors alongside wider confidence intervals and I am unable to apportion variance because it is shared between these two independent variables. These issues will make the model less interpretable and much less powerful when it comes to inference.

## Section 3

### a

In my first model I am going to omit the variable *Con10* in order to avoid violating the assumption of colinearity.

###### (8)

The summary shows us that all explanatory variables are highly significant and from the r-squared term we see that the model explains 91.2% of the variance in *Con17*. This shows that the model has exceptionally high explanatory power.

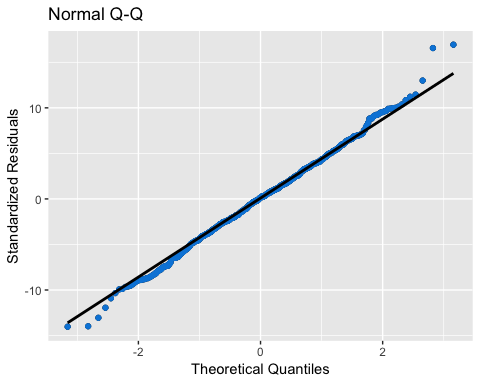
This allows us to say with confidence that for every 1 percentage point increase in the voter share for Conservatives in 2015, the vote share for Conservatives in 2017 would rise by 0.837 percentage points. The model also states that a 1 percentage point increase in the proportion of the population with white ethnicity, the vote share for Conservatives in 2017 would rise by 0.225 percentage points.

These findings agree with the scatter plots from part 2 and our hypothesis which suggested that previous Conservative vote shares had a stronger effect on the 2017 Conservative Vote Share than the proportion of the local population of white ethnicity.

### b

The assumption of normality states that the residuals in the model must be normally distributed (N~(0,). If this is satisfied then we can confirm that our model inferences and confidence intervals can be confirmed. To check, we must look at a Normal QQ-Plot which plots the residuals against a normal distribution which I will do below.

###### (9)



This graph clearly shows that *model\_1* satisfies the assumption of normality as the residuals are consistently close to the line and therefore that the calculated standard errors and their respective p-values can be trusted.

From these trusted figures I can produce a 95% confidence interval for each of the values, ‘this is an interval that contains the true value in 95% of all samples’ (Hanck et al. ([2019](#ref-hanck_econometrics_2019))).

The formula is:

1.96 is used to multiply the standard error because in a z-table the probability of getting a z-score below -1.96 is 2.5% and getting above 1.96 is also 2.5%. When added together this shows that our z-score must be between [-1.96,1.96] to achieve a 95% confidence interval.

###### (10)

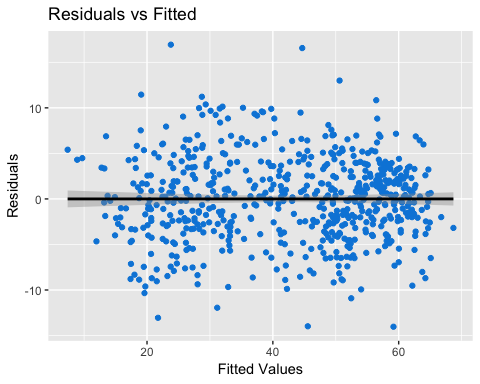
This table shows the calculated 95% confidence intervals for *model\_1*, within the *CI\_95\_Lower* and *CI\_95\_Upper* of the variables. There is a 95% chance that each interval contains their respective population mean. Due to the high significance (from the low standard errors) of the two independent variables the range between the lower and upper bound is very narrow.

## Section 4

In multivariate models, we have to satisfy ten assumptions, one of which we have just confirmed above. I will go through them now and test whether my model achieves the criteria and proves its validity.

**Linearity**: This states that all parameters I estimate in my model are linear. I can prove this by looking at the Residuals vs Fitted Plot graph. What I expect to see is a flat, horizontal line where Residuals equal zero. This will tell me that there is no pattern in the residuals and that my model is not fundamentally wrong.

###### (11)



This graph proves that my model is linear in parameters.

**Exogeneity**: This states that the regressors cannot be correlated with the error term if they are then again the model is fundamentally biased. If then it is the case that . We test this assumption with a correlation test where:

###### (12)

As we can see from this output, the p-value is equal to 1 meaning that the test has failed to reject the null hypothesis and there is no correlation between the error term and my chosen regressors.

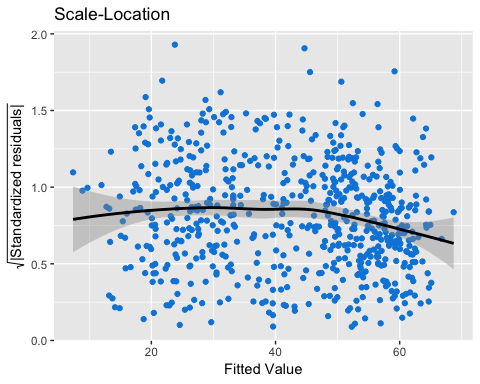
**Mean of the Residuals is zero**: The mean of the residuals must equal otherwise the model is fundamentally biased. We have already gathered good evidence to suggest my model achieves this from the Residuals vs Fitted Graph. To certify this claim I will run a t-test where:

###### (13)

As we can see from this output, the p-value equals 1 meaning that the test has failed to reject the null hypothesis and the mean of the residuals in *model\_1* is zero.

**Homoscedasticity**: Variance of the error term is constant. If this is violated then the standard errors are biased and we cannot trust their products, the p-values. To test this, we look at a Scale-Location Plot where we expect to see no discernible pattern if the model’s residuals have a constant variance. The plot itself is a visualisation of the spread of residuals over the range of predictors.

###### (14)



From the graph, *model\_1’s* residuals appear to be randomly spread over the range of residuals and so the assumption is satisfied.

**No correlation in errors**: Given a correlation between two x values, and there is no correlation in and . If this isn’t the case then standard errors are biased and p-values cannot be trusted. This is given by the equation:

This is also proved by the Residuals vs Fitted Plot I have created above when testing linearity. There is no pattern in the residuals and so we can safely safe that there is no correlation in errors.

**n >> b**: This states that the number of observations, *n*, must be larger than the number of parameters, *b*, in the model. If this is not the case then we are unable to properly observe even highly significant relationships between variables.

###### (15)

A simple calculation shows that this is satisfied.

**The Independent Variable(s) must vary**: You cannot have a homogeneous sample as you will be unable to describe the variance in the dependent variable.

I have shown the variation of my chosen independent variables in my histograms.

**Multicollinearity**: My chosen variables cannot be strongly correlated or colinear. If this is violated there is an issue of share variance, this is where you are unable to associate an effect to a specific variable. If you are unable to interpret the model then it loses its interpretability and therefore its utility.

This was an issue I originally faced with my variables as there was a 0.973 correlation between *Con15* and *Con10* which led to a 0.946 shared variance. To overcome this I omitted *Con10* from the model. Had I not actioned this then the standard errors would be much larger on both variables leading to larger confidence intervals and a confused model.

**Specification Bias**: There is an issue of whether I have chosen the correct variables or omitted a variable from the model which held high explanatory power that has been wrongly associated with another variable.

When I present my alternative model I will use an F-test to compare the two models ability to explain the *Con17* variable. The formula for the F-Test is:

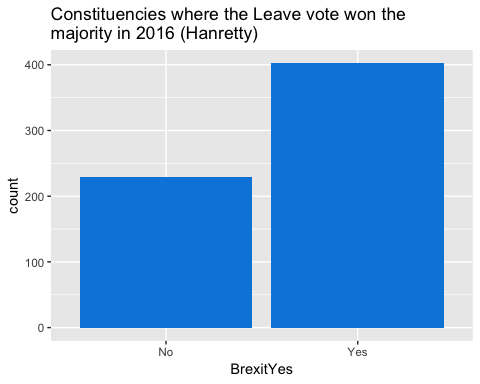
## Section 5

### a

In my second model, I propose I introduce a dummy variable *BrexitYes*, I have calculated this such that when *leaveHanretty* (Hanretty ([2017](#ref-hanretty_areal_2017))) is greater than 50% then *BrexitYes* equals 1 one and vice versa. This variable will also interact with *c11EthnicityWhite*.

I believe this model is an improvement on the first as we can see whether the Brexit verdict in each constituency effects the explanatory power of *c11EthnicityWhite* and the intercept.

###### (16)



###### (17)

###### (18)

The correlation and shared variance tables show us that our variables satisfy the multicolinearity assumption and the model as the scores are low.

###### (19)

###### (20)

The new model shows us a slightly higher r-squared value, thus means that the addition of *BrexitYes* has helped to explain the variance of our dependent term, *Con17*. All the coefficients are significant meaning that it is appropriate to interpret all of them.

This model predicts that a 1 percentage point improvement in Conservative 2015 vote share increases the 2017 Conservative vote share by 0.818 percentage points. A 1 percentage point rise in constituents with white ethnicity increases the 2017 Conservative vote share by 0.220 percentage points, if that constituency also voted to leave the EU in 2016 then this coefficient falls to 0.145. If all independent variables were zero, then we expect the Conservative 2017 vote share to be -9.262, as *Con17* they *Con17* would be 0. *Con17* would be expected to be 1.416 if the constituency’s majority vote was to leave the EU.

Like *model 1*, the 95% confidence intervals are small which is promosing for our model’s reliability.

## b

Considering the r-squared term from our first and second model we can see that the second model has more explanatory power however I will run an F-test in order to prove the superior model.

###### (21)

The result shows a DF of 2 (indicating that the more complex model has two additional parameters), and a very small p-value (2.2e-16). This means that adding the *BrexitYes* and *BrexitYes* *c11EthnicityWhite* to the model did lead to a significantly improved fit over *model 1*.

# Section 6

Both models have shown a strong and significant correlation between *Con15* and *Con17*, proving and *c11EthnicityWhite* and *Con17*, proving and that *Con15* has a larger effect on *Con17* than *c11EthnicityWhite*, proving .

I would consider using *model 2* due to its exceptionally high r-squared value with the latter model explaining 92.8% for the variance in *Con17*. My other point is that the results match those run in various other experiments, ‘the Labour party has traditionally received the lion’s share of votes from ethnic minorities’ (Martin and Khan ([2017](#ref-martin_ethnic_2017))). This suggests that our finding that constituencies with higher white populations are more likely to vote conservative is accurate. It logically follows that *c11EthnicityWhite* plays a smaller part in explaining variance than *Con15* which I have argued in my introduction.

One of the largest issues with this model is that it is built from aggregated data. This means that it can produce inaccurate inferences, constituency aggregations cannot control for individual-level determinants for example, this is an important concern known as the ecological fallacy. Statistically, a correlation tends to be larger when an association is assessed at the group level than when it is assessed at the individual level ((“Ecological Fallacy Epidemiology” [2019](#ref-noauthor_ecological_nodate))). This was first highlighted by W.S Robinson who was quoted saying his 1950s article ‘provides a definite answer as to whether ecological correlations can validly be used as substitutes for individual correlations. They cannot’ (Robinson ([1950](#ref-robinson_ecological_1950))).

What could have been beneficial to consider is whether the race of the conservative candidate, *ConPPCrace15*, would have affected the explanatory power of *c11EthnicityWhite* on *Con17*. Considering Labour’s 2017 tactic where ‘83% of Labour’s BME candidates were in seats with higher-than-average ethnic diversity’ (Martin and Khan ([2017](#ref-martin_ethnic_2017))) there is strong evidence that this variable plays a role in the verdict.

There is seemingly a relationship between homeownership and race in the UK where ‘most ethnic minority households (including White ethnic minorities) were less likely to be in homeownership than White British households’(Gov.UK ([2019b](#ref-gov.uk_home_nodate))). There is also an eleven year difference in the average age of white ethnicity and black ethnicity UK citizens (Gov.UK ([2019a](#ref-gov.uk_age_nodate))), both these variables are correlated to *Con17* says data from Akehurst ([2019](#ref-akehurst_housing_nodate)) and Curtis ([2019](#ref-curtis_how_nodate)). There could be an issue of unobserved heterogeneity which could have a major impact on results if the model doesn’t control for these variables, to find whether the *c11EthnicityWhite* variable is truly causal we must include new variables.

Overall I believe that my models hold limited use in explaining the Conservative vote share in 2017. This is due to not building the models from granular data as well as controlling for multiple more variables it is difficult to definitively and confidently derive conclusions from my models.

## Appendix

###### (1)

library(tidyverse)  
library(rlang)  
library(reshape2)  
library(tidyselect)  
library(lmtest)  
library(ggpubr)  
library(ggrepel)  
library(stargazer)  
library(e1071)  
library(broom)  
library(knitr)  
  
wd <- getwd()  
  
bes17 <- readRDS(paste0(wd, "/bes17.rds"))

###### (2)

bes17\_test <- bes17 %>%  
 dplyr:: select(Con17, Con15, Con10, c11EthnicityWhite)   
  
knitr:: kable(summary(bes17\_test))

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Con17 | Con15 | Con10 | c11EthnicityWhite |
|  | Min. : 7.301 | Min. : 4.673 | Min. : 4.40 | Min. :23.09 |
|  | 1st Qu.:29.924 | 1st Qu.:22.174 | 1st Qu.:23.65 | 1st Qu.:85.95 |
|  | Median :45.113 | Median :41.021 | Median :37.52 | Median :94.37 |
|  | Mean :42.594 | Mean :36.662 | Mean :35.69 | Mean :88.04 |
|  | 3rd Qu.:55.515 | 3rd Qu.:50.845 | 3rd Qu.:48.30 | 3rd Qu.:97.56 |
|  | Max. :69.918 | Max. :65.876 | Max. :62.80 | Max. :99.20 |
|  | NA | NA’s :1 | NA | NA |

###### (3)

histPlotR <- function(.var, mtitle){  
  
varMean <- mean(bes17[, .var], na.rm = T)  
  
varMedian <- median(bes17[, .var], na.rm = T)  
   
central\_values <- data.frame(measurement = c("mean", "median"),  
 value = c(varMean, varMedian))  
  
ggplot(data = bes17, aes\_string(x = .var)) +  
 geom\_histogram(binwidth = 2, colour = "White", fill = "#0087dc", na.rm = T) +   
 geom\_vline(data = central\_values, aes(xintercept = value, color = measurement), size = 2) +  
 scale\_color\_manual(values = c("Black", "Red"), name = NULL) +   
 ggtitle(paste0(mtitle))   
   
 }

###### (4)

bes17\_hist\_stats <- tibble(Variable = c(names(bes17\_test)),   
 Kurtosis = rep(NA, length(names(bes17\_test))),  
 Skewness = rep(NA, length(names(bes17\_test))))  
  
for (i in 1:nrow(bes17\_hist\_stats)){  
   
 bes17\_hist\_stats[i,2] <- e1071::kurtosis(bes17\_test[,i], na.rm = T, type = 2)  
   
 bes17\_hist\_stats[i,3] <- e1071::skewness(bes17\_test[,i], na.rm = T)  
   
}  
  
knitr:: kable(as.data.frame(bes17\_hist\_stats))

|  |  |  |
| --- | --- | --- |
| Variable | Kurtosis | Skewness |
| Con17 | -0.9838601 | -0.3651135 |
| Con15 | -1.2198068 | -0.2919202 |
| Con10 | -1.0184474 | -0.2804602 |
| c11EthnicityWhite | 3.8856003 | -2.0318214 |

###### (5)

scatPlotR <- function(iv, mtitle){  
   
 wrapR <- function(x, ...) {  
   
 paste(strwrap(x, ...), collapse = "\n")  
   
 }  
   
 formula <- paste0("bes17$Con17 ~ bes17$", iv)  
   
 res <- residuals(eval(call("lm", formula)))  
  
 rn <- row.names(as.data.frame(tail(sort(abs(res)), 2)))  
   
 ggplot2::ggplot(data = bes17, aes\_string(x = iv, y = "Con17")) +  
 geom\_point(colour = "#0087dc", na.rm = T) +   
 geom\_smooth(method='lm', formula = y~x, colour = "Black") +  
 geom\_text\_repel(aes(label=ifelse(row.names(bes17) %in% rn,   
 as.character(ConstituencyName),''))) +  
 coord\_cartesian(clip = 'off') +  
 ggtitle(wrapR(mtitle, 50))  
   
}

###### (6)

bes17\_cor <- cor(bes17\_test[, c("Con15", "Con10", "c11EthnicityWhite")], use = "complete.obs")  
  
knitr:: kable(round(bes17\_cor, 3))

|  |  |  |  |
| --- | --- | --- | --- |
|  | Con15 | Con10 | c11EthnicityWhite |
| Con15 | 1.000 | 0.973 | 0.233 |
| Con10 | 0.973 | 1.000 | 0.216 |
| c11EthnicityWhite | 0.233 | 0.216 | 1.000 |

###### (7)

bes17\_sv <- bes17\_cor\*bes17\_cor  
  
knitr:: kable(round(bes17\_sv, 3))

|  |  |  |  |
| --- | --- | --- | --- |
|  | Con15 | Con10 | c11EthnicityWhite |
| Con15 | 1.000 | 0.946 | 0.054 |
| Con10 | 0.946 | 1.000 | 0.047 |
| c11EthnicityWhite | 0.054 | 0.047 | 1.000 |

###### (8)

model\_1 <- lm(data = bes17\_test, formula = Con17 ~ Con15 + c11EthnicityWhite)  
  
stargazer(model\_1,  
 type = "text",  
 title = "OLS Regression Results for Model 1",  
 covariate.labels = c("Con15",  
 "c11EthnicityWhite",  
 "Intercept"),  
 summary = FALSE)

##   
## OLS Regression Results for Model 1  
## ===============================================  
## Dependent variable:   
## ---------------------------  
## Con17   
## -----------------------------------------------  
## Con15 0.837\*\*\*   
## (0.012)   
##   
## c11EthnicityWhite 0.225\*\*\*   
## (0.013)   
##   
## Intercept -7.906\*\*\*   
## (1.099)   
##   
## -----------------------------------------------  
## Observations 631   
## R2 0.912   
## Adjusted R2 0.912   
## Residual Std. Error 4.565 (df = 628)   
## F Statistic 3,259.542\*\*\* (df = 2; 628)   
## ===============================================  
## Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

###### (9)

ggplot2::qplot(sample = model\_1$residuals) +  
 stat\_qq(col = "#0087dc", na.rm = T) +  
 stat\_qq\_line(col = "Black", size = 1) +  
 xlab("Theoretical Quantiles") +  
 ylab("Standardized Residuals") +  
 ggtitle("Normal Q-Q")

###### (10)

model\_1\_ci <- summary(model\_1)$coefficients %>%  
 dplyr:: as\_tibble(rownames = c("Variables")) %>%  
 dplyr:: mutate(CI\_95\_Lower = Estimate - 1.96\*(`Std. Error`),  
 CI\_95\_Upper = Estimate + 1.96\*(`Std. Error`)) %>%  
 dplyr:: select(Variables, Estimate, CI\_95\_Lower, CI\_95\_Upper)  
  
knitr:: kable(model\_1\_ci)

|  |  |  |  |
| --- | --- | --- | --- |
| Variables | Estimate | CI\_95\_Lower | CI\_95\_Upper |
| (Intercept) | -7.9057392 | -10.0603350 | -5.7511434 |
| Con15 | 0.8374517 | 0.8147686 | 0.8601348 |
| c11EthnicityWhite | 0.2245043 | 0.1999754 | 0.2490333 |

###### (11)

ggplot2:: ggplot(model\_1, aes(model\_1$fitted.values, model\_1$residuals)) +  
 geom\_point(col = "#0087dc", na.rm = T) +  
 geom\_smooth(method='lm', formula = y~x, colour = "Black") +  
 ggtitle("Residuals vs Fitted") +  
 xlab("Fitted Values") +  
 ylab("Residuals")

###### (12)

ct <- cor.test(model\_1$fitted,model\_1$residuals)  
  
ct

##   
## Pearson's product-moment correlation  
##   
## data: model\_1$fitted and model\_1$residuals  
## t = -1.5944e-15, df = 629, p-value = 1  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## -0.078052 0.078052  
## sample estimates:  
## cor   
## -6.357257e-17

###### (13)

tt <- t.test(model\_1$residuals)  
  
tt

##   
## One Sample t-test  
##   
## data: model\_1$residuals  
## t = -2.858e-16, df = 630, p-value = 1  
## alternative hypothesis: true mean is not equal to 0  
## 95 percent confidence interval:  
## -0.3563364 0.3563364  
## sample estimates:  
## mean of x   
## -5.186043e-17

###### (14)

ggplot2:: ggplot(model\_1, aes(model\_1$fitted.values, sqrt(abs(.stdresid)))) +  
 geom\_point(col = "#0087dc", na.rm=TRUE) +  
 stat\_smooth(method="loess", col = "Black" ,na.rm = TRUE) +   
 ggtitle("Scale-Location") +  
 xlab("Fitted Value") +  
 ylab(expression(sqrt("|Standardized residuals|")))

###### (15)

nrow(bes17\_test) > (ncol(bes17\_test)-1)

## [1] TRUE

###### (16)

bes17 <- bes17 %>%  
 dplyr:: mutate(BrexitYes = ifelse(leaveHanretty > 50, "Yes","No"))  
  
 wrapR <- function(x, ...) {  
   
 paste(strwrap(x, ...), collapse = "\n")  
   
 }  
  
ggplot2:: ggplot(bes17, aes(BrexitYes)) +  
 geom\_bar(fill = "#0087dc") +  
 ggtitle(wrapR("Constituencies where the Leave vote won the majority in 2016 (Hanretty)", 50))

bes17 <- bes17 %>%  
 dplyr:: mutate(BrexitYes = ifelse(BrexitYes =="Yes",1,0))

###### (17)

#dftst <- dplyr:: select(bes17, Con15, c11EthnicityWhite, BrexitYes)  
  
bes17\_cor2 <- cor(bes17[,c("Con15", "c11EthnicityWhite", "BrexitYes"),], use = "complete.obs")  
  
knitr:: kable(round(bes17\_cor2, 3))

|  |  |  |  |
| --- | --- | --- | --- |
|  | Con15 | c11EthnicityWhite | BrexitYes |
| Con15 | 1.000 | 0.233 | 0.242 |
| c11EthnicityWhite | 0.233 | 1.000 | 0.287 |
| BrexitYes | 0.242 | 0.287 | 1.000 |

###### (18)

bes17\_sv2 <- bes17\_cor2\*bes17\_cor2  
  
knitr:: kable(round(bes17\_sv2, 3))

|  |  |  |  |
| --- | --- | --- | --- |
|  | Con15 | c11EthnicityWhite | BrexitYes |
| Con15 | 1.000 | 0.054 | 0.059 |
| c11EthnicityWhite | 0.054 | 1.000 | 0.083 |
| BrexitYes | 0.059 | 0.083 | 1.000 |

###### (19)

model\_2 <- lm(data = bes17, formula = Con17 ~ Con15 + c11EthnicityWhite + BrexitYes + BrexitYes\*c11EthnicityWhite)  
  
stargazer(model\_2,  
 type = "text",  
 title = "OLS Regression Results for Model 2",  
 covariate.labels = c("Con15",  
 "c11EthnicityWhite",  
 "BrexitYes",  
 "BrexitYes\*c11EthnicityWhite",  
 "Intercept"))

##   
## OLS Regression Results for Model 2  
## =======================================================  
## Dependent variable:   
## ---------------------------  
## Con17   
## -------------------------------------------------------  
## Con15 0.818\*\*\*   
## (0.011)   
##   
## c11EthnicityWhite 0.220\*\*\*   
## (0.015)   
##   
## BrexitYes 10.678\*\*\*   
## (2.116)   
##   
## BrexitYes\*c11EthnicityWhite -0.075\*\*\*   
## (0.024)   
##   
## Intercept -9.262\*\*\*   
## (1.261)   
##   
## -------------------------------------------------------  
## Observations 631   
## R2 0.928   
## Adjusted R2 0.928   
## Residual Std. Error 4.137 (df = 626)   
## F Statistic 2,019.563\*\*\* (df = 4; 626)   
## =======================================================  
## Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

###### (20)

model\_2\_ci <- summary(model\_2)$coefficients %>%  
 dplyr:: as\_tibble(rownames = c("Variables")) %>%  
 dplyr:: mutate(CI\_95\_Lower = Estimate - 1.96\*(`Std. Error`),  
 CI\_95\_Upper = Estimate + 1.96\*(`Std. Error`)) %>%  
 dplyr:: select(Variables, Estimate, CI\_95\_Lower, CI\_95\_Upper)  
  
knitr:: kable(model\_2\_ci)

|  |  |  |  |
| --- | --- | --- | --- |
| Variables | Estimate | CI\_95\_Lower | CI\_95\_Upper |
| (Intercept) | -9.2622652 | -11.7341276 | -6.7904027 |
| Con15 | 0.8181908 | 0.7971486 | 0.8392330 |
| c11EthnicityWhite | 0.2199261 | 0.1911211 | 0.2487311 |
| BrexitYes | 10.6775552 | 6.5298391 | 14.8252713 |
| c11EthnicityWhite:BrexitYes | -0.0747050 | -0.1213832 | -0.0280268 |

###### (21)

anova(model\_1, model\_2)

## Analysis of Variance Table  
##   
## Model 1: Con17 ~ Con15 + c11EthnicityWhite  
## Model 2: Con17 ~ Con15 + c11EthnicityWhite + BrexitYes + BrexitYes \* c11EthnicityWhite  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 628 13090   
## 2 626 10714 2 2375.9 69.413 < 2.2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

## Bibliography

Akehurst, Steve. 2019. “Housing and the 2017 Election: What the Numbers Say Shelter.” Accessed November 23. <https://blog.shelter.org.uk/2017/06/housing-and-the-2017-election-what-the-numbers-say/>.

Alec, Tyson. 2016. “Behind Trump’s Victory: Divisions by Race, Gender, Education,” November. <https://www.pewresearch.org/fact-tank/2016/11/09/behind-trumps-victory-divisions-by-race-gender-education/>.

Baker, Carl, Oliver Hawkins, Lukas Audickas, Alex Bate, Richard Cracknell, Vyara Apostolova, Noel Dempsey, Roderick McInnes, Tom Rutherford, and Elise Uberoi. 2019. “General Election 2017: Full Results and Analysis,” January. <https://researchbriefings.parliament.uk/ResearchBriefing/Summary/CBP-7979>.

Cowley, Philip, and Dennis Kavanagh. 2016. *The British General Election of 2015*. British General Election Series. Palgrave Macmillan UK. doi:[10.1057/9781137366115](https://doi.org/10.1057/9781137366115).

———. 2018. *The British General Election of 2017*. British General Election Series. Palgrave Macmillan. doi:[10.1007/978-3-319-95936-8](https://doi.org/10.1007/978-3-319-95936-8).

Cracknell, Richard, Nick Duckworth, Oliver Hawkins, and John Wood. 2014. “Census 2011 Constituency Results.” House of Commons.

Curtis, Chris. 2019. “How Britain Voted at the 2017 General Election YouGov.” Accessed November 23. <https://yougov.co.uk/topics/politics/articles-reports/2017/06/13/how-britain-voted-2017-general-election>.

“Ecological Fallacy Epidemiology.” 2019. *Encyclopedia Britannica*. Accessed November 23. <https://www.britannica.com/science/ecological-fallacy>.

Gov.UK. 2019a. “Age Groups.” Accessed November 23. <https://www.ethnicity-facts-figures.service.gov.uk/uk-population-by-ethnicity/demographics/age-groups/latest#average-age-by-ethnicity>.

———. 2019b. “Home Ownership.” Accessed November 23. <https://www.ethnicity-facts-figures.service.gov.uk/housing/owning-and-renting/home-ownership/latest#by-ethnicity>.

Hanck, Christoph, Martin Arnold, Alexander Gerber, and Martin Schmelzer. 2019. “Econometrics with R.” *Introduction to Econometrics with R*. <https://www.econometrics-with-r.org/>.

Hanretty, Chris. 2017. “Areal Interpolation and the UK’s Referendum on EU Membership.” *Journal of Elections, Public Opinion and Parties* 27 (4): 466–83. doi:[10.1080/17457289.2017.1287081](https://doi.org/10.1080/17457289.2017.1287081).

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2013. *An Introduction to Statistical Learning*. Vol. 103. Springer Texts in Statistics. New York, NY: Springer New York. doi:[10.1007/978-1-4614-7138-7](https://doi.org/10.1007/978-1-4614-7138-7).

Khan, Omar, Kjartan Páll Sveinsson, and Runnymede Trust. 2015. *Race and Elections*.

Martin, Nicole, and Omar Khan. 2017. “Ethnic Minorities at the 2017 British General Election,” 4.

McNeese, Bill. 2008. “Are the Skewness and Kurtosis Useful Statistics?” <https://www.spcforexcel.com/knowledge/basic-statistics/are-skewness-and-kurtosis-useful-statistics#skewness>.

Robinson, W. S. 1950. “Ecological Correlations and the Behavior of Individuals.” *American Sociological Review* 15 (3): 351–57. doi:[10.2307/2087176](https://doi.org/10.2307/2087176).

tutor2u. 2019. “Safe Seats.” Text/html. *Tutor2u*. <https://www.tutor2u.net/politics/reference/safe-seats>.

Westfall, Peter H. 2014. “Kurtosis as Peakedness, 1905 – 2014. R.I.P.” *The American Statistician* 68 (3): 191–95. doi:[10.1080/00031305.2014.917055](https://doi.org/10.1080/00031305.2014.917055).