MATH158 Final Project: Predicting United States GDP and Political Affiliation Using Energy Data

Kat Gelsey, Sarah Kim, and Paul McKinley

12/10/2021

ABSTRACT: This project explores how one sector of the U.S. economy can impact broader economic and societal outcomes. With an increasing focus on the energy production and consumption, as well heightened support for a global transition to renewable energy sources, we focus on how the U.S. energy sector affects high-level variables. We use the source of electricity produced in a state and that state's total energy consumption to predict total state GDP, as well as political affiliation. Here we construct three models to explore these questions. First, we find that total energy consumption and a state's total GDP are highly correlated, suggesting unsurprisingly that wealthier states consume more energy. Second, we construct a model with GDP as a response and electricity production source and total energy consumption as predictors, finding that solar and nuclear are significant predictors, along with total consumption. Finally, we construct a model with percentage of votes in a state going to Republican or Democrat candiates in the U.S. 2020 presidential election as a response, finding that nuclear, coal, natural gas, petroleum, geothermal and wind are significant predictors. We generally conclude that energy can have a moderate level of predictive impact on broader economic and political shifts; however, our modeling is limited by lack of information, collinearity, and nonuniform distributions of electricity generation type in U.S. states.

I. INTRODUCTION

The objective of this project is to analyze how one sector of the U.S. economy (i.e. energy) can influence broader societal outcomes, such as economic growth and elections. In this case, we are interested in building models based on energy supply (for electricity) in each state, to see what role the energy sector may have on a broader outcome. This report addresses the following questions:

(1) Is electricity production type correlated with total GDP per state?
(2) Which sources of electricity production is most significant in predicting GDP?
(3) In addition to energy production, is total energy consumption a significant predictor for total GDP per state?
(4) Is electricity production type correlated with state-level political affiliation?

Our project is broken into three components. First, we analyze total GDP with total energy consumption by state as a predictor to see if total energy consumption is relevant to total GDP.

Then, we aim to investigate the relationship between energy production source and total energy consumption against and total GDP. This allows us to compare both energy production and comsumption in the context of our societal variables of interest. Our energy data comes from the National Energy Institute, providing the percentage of electricity that is produced in a given state by each source, such as solar, wind, coal, natural gas, etc. Our GDP data comes from Statistia Research Department.

Finally, we are interested in the relationship between energy production and political affiliation. In this case, we use the percentage of voters who voted Republican vs. Democrat in the 2020 presidential race as a response, and electric energy production types and total energy consumption as predictors.

II. DATA — Describe details about how the data set was collected and the variables in the data set.

Our data consists of four datasets, merged by U.S. state: (1) We use total GDP by state in 2020 (in billions of US dollars) with data collected by the Statista Research Department. This is our primary societal-level economic response variable. It is worth noting that GDP is only one metric of general economic trends, comprised of multiple sectors beyond our focus of energy. Additionally, this project does not adjust for the impact on GDP caused by COVID-19 during the 2020 fiscal year. https://www.statista.com/statistics/248023/us-gross-domestic-product-gdp-by-state/

(2) We use electricity production data from the National Energy Institute (NEI) in 2020 as our primary source of predictor variables. Data is organized by percent of total electricity produced in the fifty U.S. states and the District of Columbia by a subset of electricity production types. In this case we use nuclear, solar, coal, natural gas, hydroelectric, petroleum, wind, geothermal, and biomass/other. We note that electricity production is only one component of the energy sector, and that in many states, not all of these sectors have significant percentages of the total electricity production. However, electricity is of greater proximity to residences and businesses, and thus could have a higher impact on outside economic outcomes. https://www.nei.org/resources/statistics/state-electricity-generation-fuel-shares

(3) In addition to energy production, we also explore total energy consumption (in British thermal units) by state in 2018 (the latest this data was avaialable). Although we do not have data for consumption by electricity sector, as in the production dataset, we are still able to explore the distinction between production and consumption. https://neo.ne.gov/programs/stats/inf/120.htm

(4) To measure political affiliation in the U.S. with one response variable, we use data from the U.S. 2020 presidential election. Here we use data collected by Cook Political Report, exploring percentages of votes in a given state that went to the Republican or Democratic candidate. We do not explore independent candidates because of the small proportion of votes they comprise. We note that percentage of votes for each candidate is only one representation of political affiliation. https://www.census.gov/newsroom/press-releases/2021/2020-presidential-election-voting-and-registration-tables-now-available.html

To efficiently conduct our analysis, we merged these data sets into a single data frame called "fulldata", which we invoke throughout our analysis:

MAKING THE FULL DATASET

```
#require(dplyr)
require(tidyverse)
```

```
## Loading required package: tidyverse

## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.5     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.0.1     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
require(MASS)
```

```
## Loading required package: MASS

##
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':
##
##     select
```

```
require(faraway)
```

```
## Loading required package: faraway
```

```
require(ggplot2)
require(ggpmisc)
```

```
## Loading required package: ggpmisc

## Loading required package: ggpp

##
## Attaching package: 'ggpp'

## The following object is masked from 'package:ggplot2':
##
##     annotate

## Warning in .recacheSubclasses(def@className, def, env): undefined subclass
## "numericVector" of class "Mnumeric"; definition not updated
```

```
require(pls)
```

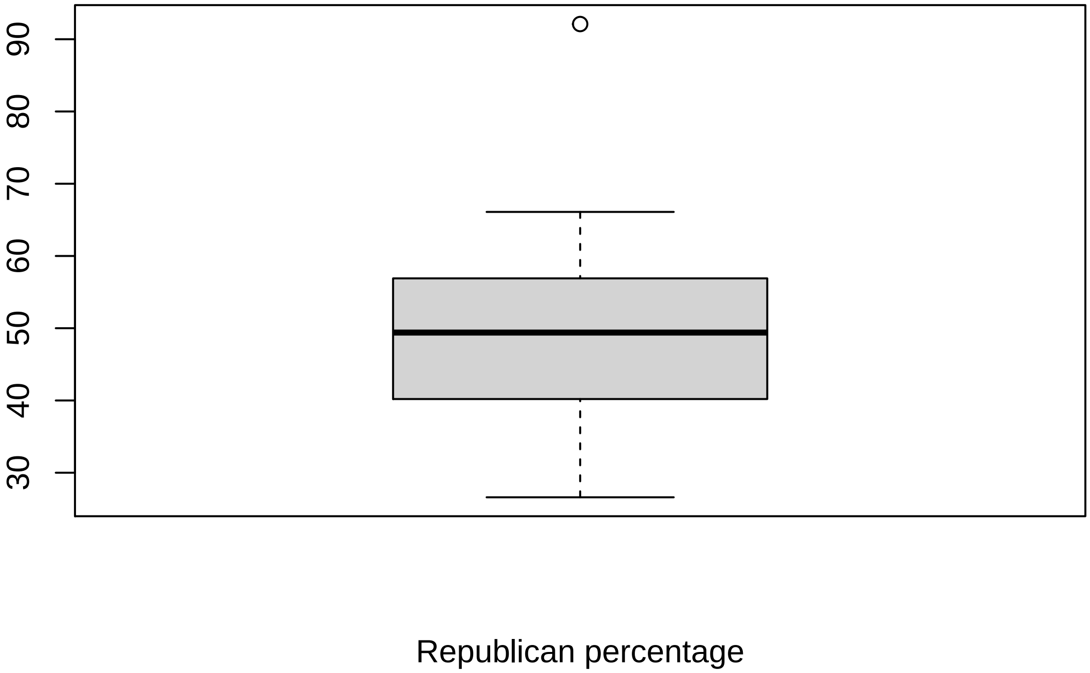
```
## Loading required package: pls

##
## Attaching package: 'pls'

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

```
fulldata <-read.csv("Project Data/masterdata.csv")
```

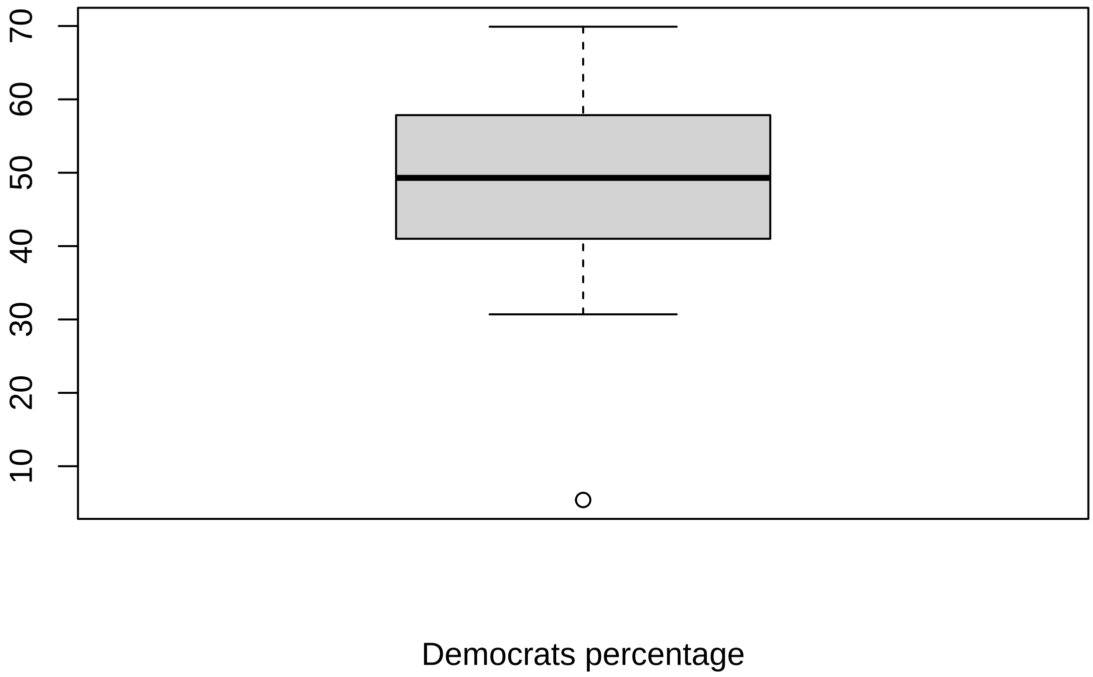
- insert basic observations about boxplot of the data *

Response variables: total GDP & Political Affiliation:

```
boxplot(fulldata$dem_percent, xlab = "Republican percentage", main = "")
```

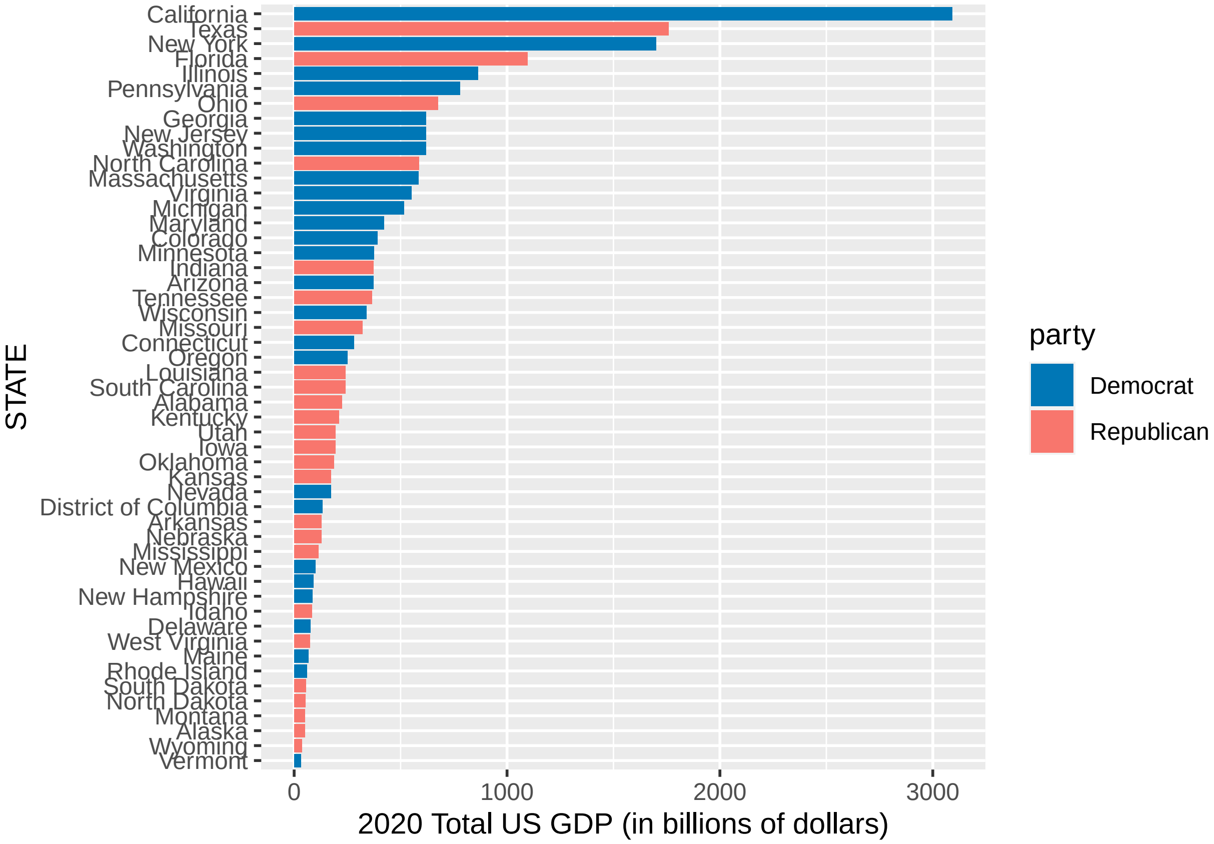


```
boxplot(fulldata$rep_percent, xlab = "Democrats percentage", main = "")
```



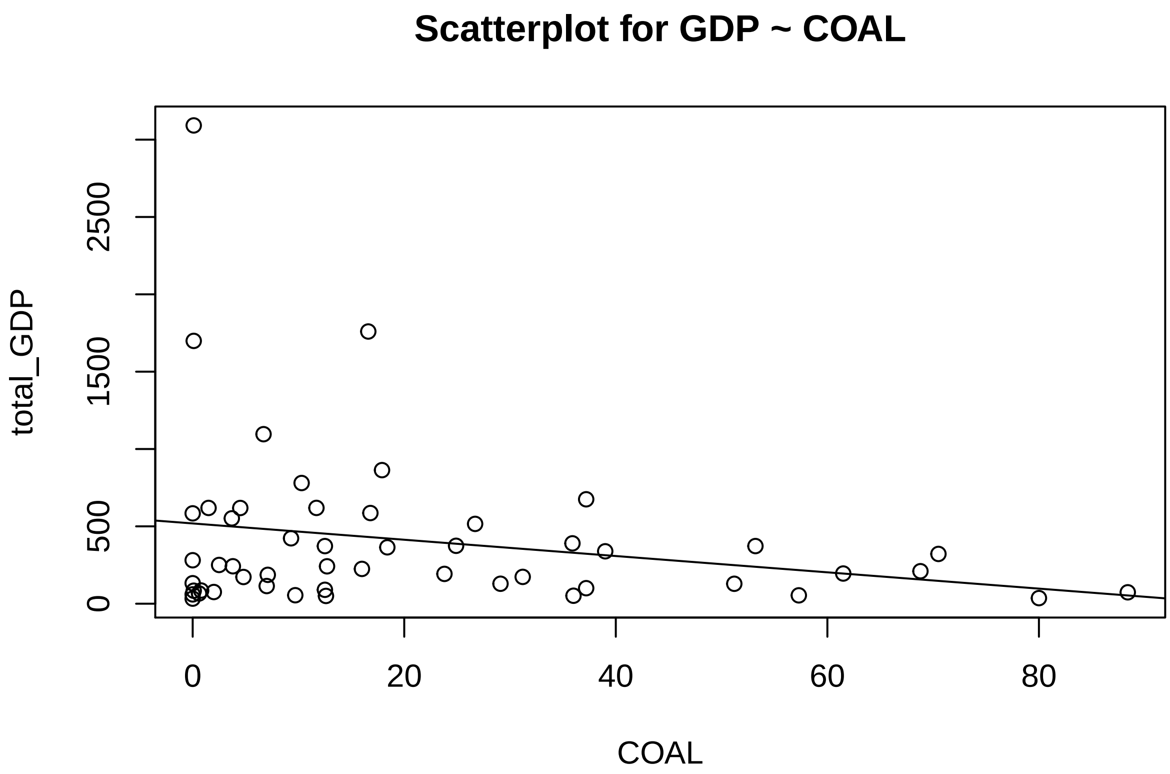
We might want to consider doing a log transformation of our response variable “total GDP” because there seems to be an extreme outlier for the total GDP (i.e California).

```
#plotting total US GDP
ggplot(data = fulldata, mapping = aes(x = reorder(STATE, total_GDP), y=total_GDP, fill=party)) +
  scale_fill_manual(values=c("#0077b6", "#F8766D"))+
  geom_bar(stat = "identity") + coord_flip()+
  labs(x="STATE", y="2020 Total US GDP (in billions of dollars)")
```



From the first look, it seems like our response variables and the explanatory variables has some correlations. *Explanatory variables ~ total GDP*:

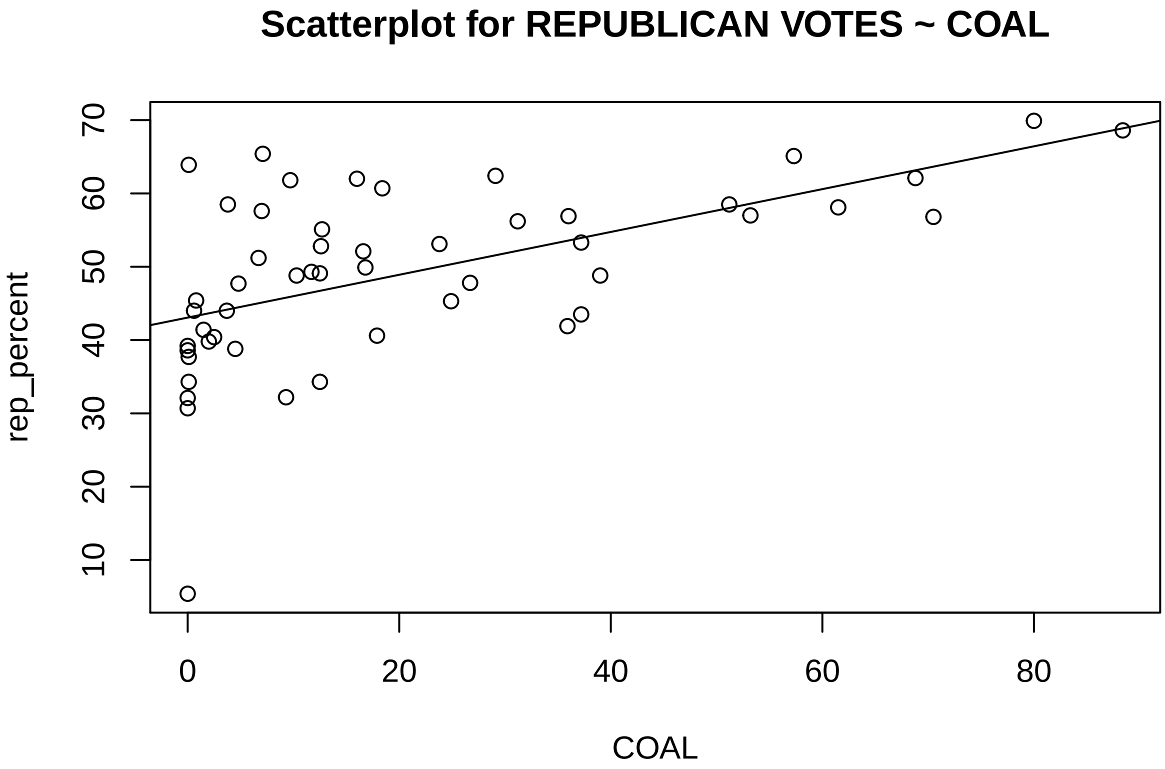
```r
# response variable GDP ~ some numerical variables of interest
plot(total_GDP ~ COAL, fulldata, main = "Scatterplot for GDP ~ COAL")
abline(lm(total_GDP ~ COAL, fulldata))
```



```
# plot(total_GDP ~ PETROLEUM, fulldata, main = "Scatterplot for GDP ~ PETROLEUM")
# abline(lm(total_GDP ~ PETROLEUM, fulldata))
#
# plot(total_GDP ~ NUCLEAR, fulldata, main = "Scatterplot for GDP ~ NUCLEAR")
# abline(lm(total_GDP ~ PETROLEUM, fulldata))
#
# plot(total_GDP ~ WIND, fulldata, main = "Scatterplot for GDP ~ WIND")
# abline(lm(total_GDP ~ PETROLEUM, fulldata))
```

Explanatory variables ~ Political Affiliation:

```
# response variable GDP ~ some numerical variables of interest
plot(rep_percent ~ COAL, fulldata, main = "Scatterplot for REPUBLICAN VOTES ~ COAL")
abline(lm(rep_percent ~ COAL, fulldata))
```



```
#plot(rep_percent ~ PETROLEUM, fulldata, main = "Scatterplot for REPUBLICAN VOTES ~ PETROLEUM")
#abline(lm(rep_percent ~ PETROLEUM, fulldata))

#plot(rep_percent ~ NUCLEAR, fulldata, main = "Scatterplot for REPUBLICAN VOTES ~ NUCLEAR")
#abline(lm(rep_percent ~ PETROLEUM, fulldata))

#plot(rep_percent ~ WIND, fulldata, main = "Scatterplot for REPUBLICAN VOTES ~ WIND")
#abline(lm(rep_percent ~ PETROLEUM, fulldata))
```

III. ANALYSIS We are interested to see how energy, a specific sector of an economy, can affect the total GDP. Especially since From our two preliminary models, we have concluded that total energy consumption and electricity production predictors are significant predictors of the US total GDP.

A. Preliminary Model 1: GDP Growth ~ Total Energy Consumption by state Our first model looks at how total energy consumption can influence the total GDP. We wanted to look at the total energy consumption per state to see how much of energy demand we have for each state. After we create the model and look at the summary, we see that a state's energy consumption appears to be an extremely significant predictor of the state's GDP. With this predictor alone we get an R^2 value of 0.64, and the p-value for the energy consumption predictor is 2.31e-12.

```
GDPcon <- lm(total_GDP ~ consumption_Tbtu, fulldata)
summary(GDPcon)
```

```
##
## Call:
## lm(formula = total_GDP ~ consumption_Tbtu, data = fulldata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -931.41 -104.66  -46.57   58.42 1570.97
##
```

```
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)    
## (Intercept)      39.25570   60.38354   0.650    0.519    
## consumption_Tbtu  0.18598    0.02005   9.274 2.31e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 324.6 on 49 degrees of freedom
## Multiple R-squared:  0.6371, Adjusted R-squared:  0.6297 
## F-statistic: 86.01 on 1 and 49 DF,  p-value: 2.31e-12
```

##BOXCOX After creating the model, our first step is to determine whether the response (total GDP) needs to be transformed. The boxcox() function plot tells us that transformation is necessary, since the confidence interval for λ is centered at 0.2 and does not contain 1. For convenience and consistency, we use $\lambda = 0$ and log-transform the response.

```
# boxcox(GDPcon, lambda=seq(-0.25,0.75,by=0.05),plotit=T)
# GDPcon_trans <- lm(log(total_GDP) ~ consumption_Tbtu, fulldata)
# sumary(GDPcon_trans)
```

##DIAGNOSTICS Even for the transformed data, we see that the plot of the residuals is not random, with the largest fitted values having extremely high or low residuals. To check for normality, we look at a QQ plot of the residuals. There is some long-tailed distribution, but not as bad as when we plotted the QQ plot with the untransformed data, which had a long-tailed distribution. The Shapiro-Wilk test demonstrates that the data is now normal, in contrast with the test's highly significant p-value using the untransformed data. (Faraway 75-81)

```
#plot(GDPcon_trans)
#shapiro.test(residuals(GDPcon_trans))
```

Obsservation 33 (New York) and 44 (Texas) are leverage points, influential points, and outliers.

Removing New York from the model significantly improves the R^2 value for the model and drastically increases the significance of the consumption_Tbtu predictor; the same observations are true when we remove Texas.

```
fulldata[33,]#new york
```

```
##     X    STATE total_GDP NUCLEAR COAL NATURAL.GAS PETROLEUM HYDRO GEOTHERMAL
## 33 33 New York   1699.04    29.1  0.1        40.1       0.2  23.6          0
##    SOLARPV WIND BIOMASS_OTHER    party    dem_votes    rep_votes other_votes
## 33     0.8  3.8           2.3 Democrat 5,244,886.00 3,251,997.00  119,978.00
##    dem_percent rep_percent other_percent Rank consumption_Tbtu  index
## 33        60.9        37.7           1.4   45           3854.2 0.2703
```

```
GDPcon_trans_33 <- lm((total_GDP)^(0.2) ~ consumption_Tbtu, fulldata[-33,])
sumary(GDPcon_trans_33)
```

```
##                    Estimate Std. Error t value  Pr(>|t|)
## (Intercept)      2.5958e+00 7.6339e-02 34.0040 < 2.2e-16
## consumption_Tbtu 2.1471e-04 2.5516e-05  8.4147 5.239e-11
## 
## n = 50, p = 2, Residual SE = 0.41014, R-Squared = 0.6
```

```
fulldata[44,]#texas
```

```
##     X STATE total_GDP NUCLEAR COAL NATURAL.GAS PETROLEUM HYDRO GEOTHERMAL
## 44 44 Texas   1759.73     8.7 16.6        52.7         0   0.4          0
##    SOLARPV WIND BIOMASS_OTHER      party    dem_votes    rep_votes other_votes
## 44     1.7 19.6           0.4 Republican 5,259,126.00 5,890,347.00  165,583.00
##    dem_percent rep_percent other_percent Rank consumption_Tbtu index
## 44        46.5        52.1           1.5   51          14258.8     1
```

```
GDPcon_trans_44 <- lm((total_GDP)^(0.2) ~ consumption_Tbtu, fulldata[-44,])
sumary(GDPcon_trans_44)
```

```
##                   Estimate Std. Error t value  Pr(>|t|)
## (Intercept)      2.3646e+00 6.9383e-02  34.081 < 2.2e-16
## consumption_Tbtu 3.7318e-04 3.0481e-05  12.243 2.249e-16
##
## n = 50, p = 2, Residual SE = 0.31708, R-Squared = 0.76
```

(Since we have only one predictor, it is irrelevant to consider collinearity, stepwise/backwards regression, or PCR.)

Preliminary Model 2: total GDP ~ Energy production source by type for each state

Our second model looks at how energy production by type could be a predictor of gross domestic product. We again transformed our response variable by taking its log. With our second model, we want to find which energy production by type are significant predictors for state GDP. Setting GDP as response and energy production types as predictors, we were able to conclude that solar, nuclear, and natural gas are the significant predictors. However, the R^2 value is significantly low.

```
# linear model 2
GDPpro_log <- lm(log(total_GDP) ~ NUCLEAR + COAL+ NATURAL.GAS + PETROLEUM + HYDRO + GEOTHERMAL + SOLARPV
# step(GDPpro_log)
GDPpro_log_step <- lm(formula = log(total_GDP) ~ NUCLEAR + COAL + NATURAL.GAS +
    PETROLEUM + HYDRO + SOLARPV + WIND, data = fulldata)
summary(GDPpro_log_step)
```

```
##
## Call:
## lm(formula = log(total_GDP) ~ NUCLEAR + COAL + NATURAL.GAS +
##     PETROLEUM + HYDRO + SOLARPV + WIND, data = fulldata)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.61570 -0.65579 -0.09938  0.57935  1.85317
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.78257    2.86231  -0.273  0.78585
## NUCLEAR      0.07985    0.02950   2.707  0.00971 **
## COAL         0.05325    0.02820   1.888  0.06576 .
## NATURAL.GAS  0.06411    0.02961   2.166  0.03594 *
## PETROLEUM    0.04587    0.03306   1.387  0.17245
## HYDRO        0.05250    0.03131   1.677  0.10087
## SOLARPV      0.16833    0.06460   2.606  0.01255 *
## WIND         0.05804    0.03102   1.871  0.06812 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9097 on 43 degrees of freedom
## Multiple R-squared:  0.3736, Adjusted R-squared:  0.2717
## F-statistic: 3.665 on 7 and 43 DF,  p-value: 0.003455
```

Full Model 1: total GDP ~ Energy production source by type + Total energy consumption

Moving forward, we wanted to construct a model that predicts GDP with both total energy consumption and energy production source types as predictors. We then examined if energy production alone is enough to predict GDP by running an ANOVA analysis.

We again log-transform the response (GDP_nt –> GDPfull) after viewing the boxcox() plot.

```
GDP_nt <-lm(total_GDP ~ NUCLEAR + COAL+ NATURAL.GAS + PETROLEUM + HYDRO + GEOTHERMAL + SOLARPV + WIND +
# bc <- boxcox(GDP_nt, lambda=seq(-0.5,0.5,by=0.05), plotit = T)
GDPfull <-lm(log(total_GDP) ~ NUCLEAR + COAL+ NATURAL.GAS + PETROLEUM + HYDRO + GEOTHERMAL + SOLARPV +
anova(GDPfull, GDPpro_log)
```

```
## Analysis of Variance Table
##
## Model 1: log(total_GDP) ~ NUCLEAR + COAL + NATURAL.GAS + PETROLEUM + HYDRO +
##     GEOTHERMAL + SOLARPV + WIND + BIOMASS_OTHER + consumption_Tbtu
## Model 2: log(total_GDP) ~ NUCLEAR + COAL + NATURAL.GAS + PETROLEUM + HYDRO +
##     GEOTHERMAL + SOLARPV + WIND + BIOMASS_OTHER
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     40 17.885
## 2     41 35.146 -1    -17.26 38.602 2.371e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Since the p-value for the ANOVA is smaller than 0.05, we can conclude that the reduced model does not capture as much as the larger model; in other words, the bigger model with both consumption and production as predictors is preferred. Hence, we decided to continue with the bigger model 'GDPfull' and run stepwise regression analysis to see which explanatory variables are significant.

```
# step(GDPfull)
GDPfull_step <- lm(formula = log(total_GDP) ~ NUCLEAR + SOLARPV + consumption_Tbtu,
    data = fulldata)
summary(GDPfull_step)
```

```
##
## Call:
## lm(formula = log(total_GDP) ~ NUCLEAR + SOLARPV + consumption_Tbtu,
##     data = fulldata)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.51870 -0.33124  0.01246  0.38456  1.24024
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      4.382e+00  1.552e-01  28.245  < 2e-16 ***
## NUCLEAR          2.103e-02  5.304e-03   3.965 0.000248 ***
## SOLARPV          5.995e-02  2.702e-02   2.219 0.031360 *
## consumption_Tbtu 3.033e-04  4.078e-05   7.437 1.78e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6466 on 47 degrees of freedom
## Multiple R-squared:  0.6542, Adjusted R-squared:  0.6321
## F-statistic: 29.64 on 3 and 47 DF,  p-value: 6.609e-11
```

Our stepwise regression model indicates that nuclear energy production, solar energy production, and total energy consumption are the best predictors. It suggests that an increase in total energy consumption leads to an increase in the total GDP, which intuitively makes sense. Yet, what really caught our attention is the fact that for every one percent increase in net electricity produced by nuclear and/or solar energy, there is an increase in the total GDP. In the context of our electricity production dataset, which measures the percentage of net electricity generated with the given production method, our linear model suggests that shifting towards nuclear and solar energy production type generated a greater increase in the total GDP produced by state. However, there are some caveats to consider (see Discussion).

```
#anova
anova(GDPfull, GDPfull_step)#reduced is fine
```

```
## Analysis of Variance Table
##
## Model 1: log(total_GDP) ~ NUCLEAR + COAL + NATURAL.GAS + PETROLEUM + HYDRO +
##     GEOTHERMAL + SOLARPV + WIND + BIOMASS_OTHER + consumption_Tbtu
## Model 2: log(total_GDP) ~ NUCLEAR + SOLARPV + consumption_Tbtu
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     40 17.885
## 2     47 19.647 -7   -1.7619 0.5629 0.7813
```

Comparing our models–one with all predictors and the reduced model–with ANOVA, we can conclude that the reduced model does as well as the bigger model. We can conclude that nuclear energy, solar energy, and total energy consumption are indeed the most important predictors for this model. This is further confirmed by constructing 95% confidence intervals for the stepped model:

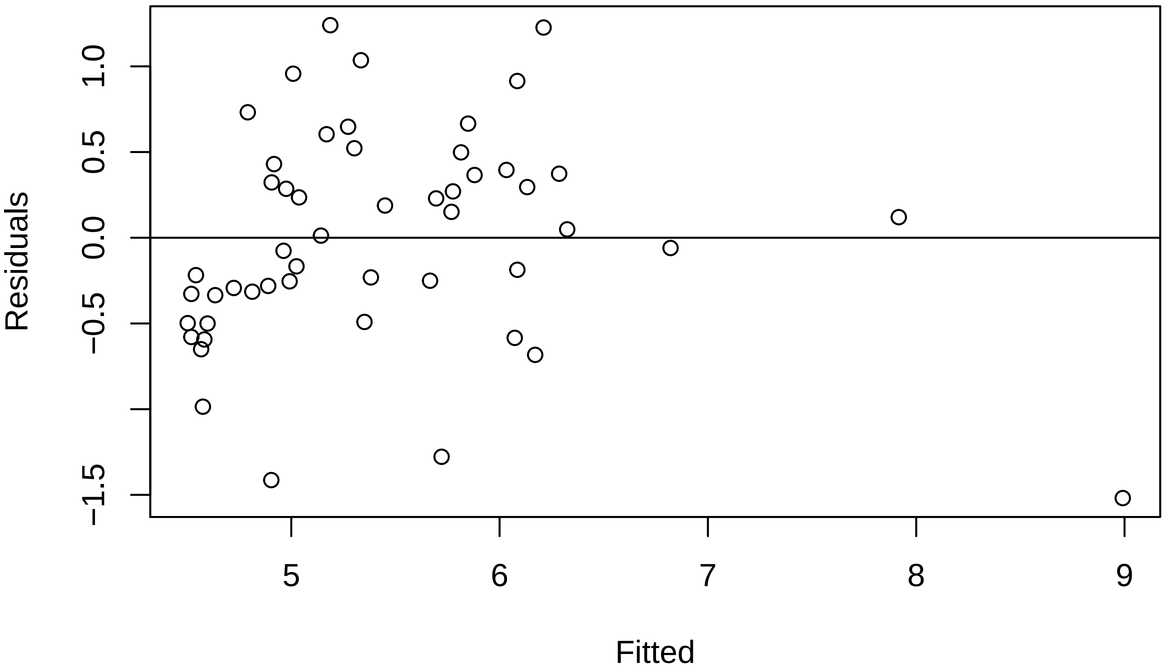
```
# 95% CI
confint(GDPfull_step)#all significant
```

```
##                         2.5 %      97.5 %
## (Intercept)      4.0702794937 4.694550246
## NUCLEAR          0.0103610678 0.031699980
## SOLARPV          0.0055961922 0.114297385
## consumption_Tbtu 0.0002212351 0.000385313
```

We see that 0 is not contained in the range of any of the confidence intervals for the selected predictors, so we can conclude that the coefficients of these predictors differ significantly from 0.

#DIAGNOSTICS We examine the distribution of residuals and unusual observations in the model *after* stepwise regression.

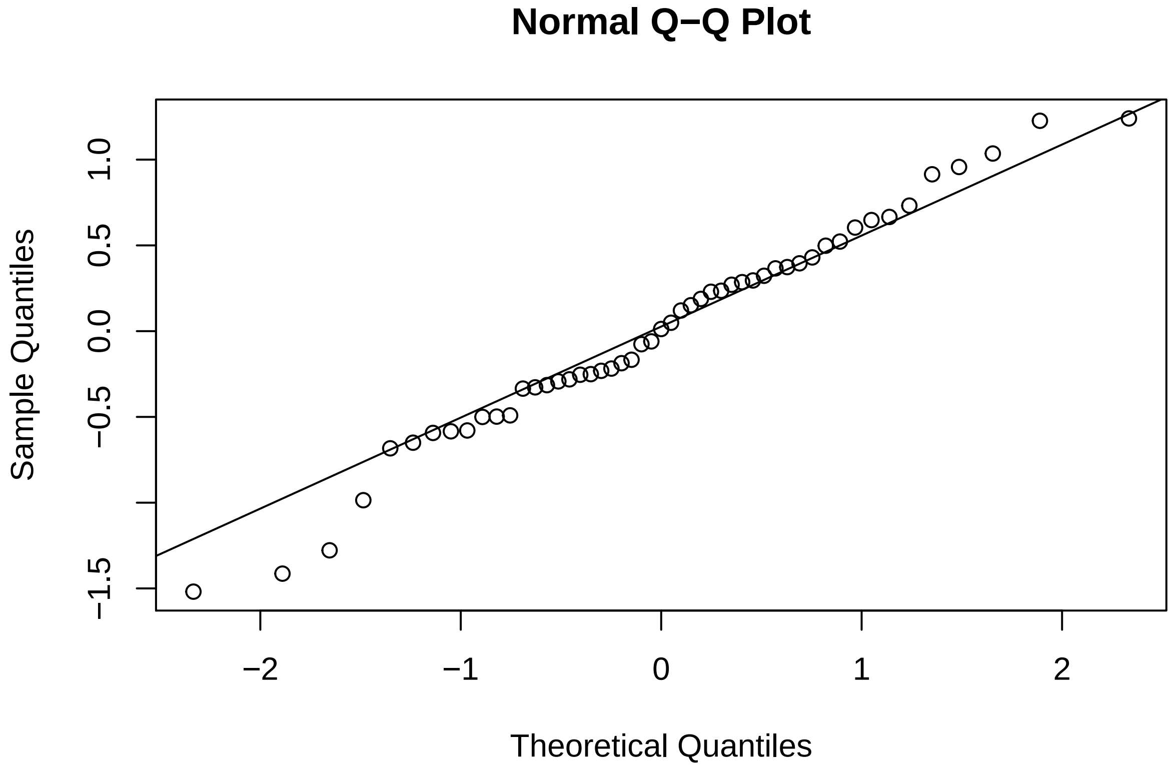
```
plot(residuals(GDPfull_step) ~ fitted(GDPfull_step), xlab="Fitted",ylab="Residuals")
abline(h=0)
```



We can see there is an underlying non-symmetrical structure of the residuals when we plot them against the fitted values. The model, therefore, does not satisfy constant variance. We discuss potential workarounds to this issue later (see "Mitigating Problems with the Error").

##Check for Normality

```
qqnorm(residuals(GDPfull_step))
qqline(residuals(GDPfull_step))
```



```
shapiro.test(residuals(GDPfull_step))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(GDPfull_step)
## W = 0.9821, p-value = 0.6308
```

Even though we see some evidence of a long-tailed distribution based on the QQ plot, the Shapiro-Wilk test p-value is greater than 0.05; therefore, the residuals are normally distributed.

#Correlation between Predictors and Collinearity For our energy production and consumption data, we are interested in the potential correlations between predictors.

We construct a model matrix on the full model to compute correlation coefficients between predictors. We can use this correlation matrix as a reference for other models that use a different response (as in the case for our political affiliation model) or have a reduced number of predictors.

```
GDPfullMat<-model.matrix(GDPfull)
(cormat <- cor(GDPfullMat[-1,-1])) #get rid of intercept
```

```
##                       NUCLEAR        COAL  NATURAL.GAS   PETROLEUM       HYDRO
## NUCLEAR           1.000000000 -0.24241725 -0.002309012 -0.14960697 -0.27601173
## COAL             -0.242417247  1.00000000 -0.496646405 -0.06454496 -0.24816204
## NATURAL.GAS      -0.002309012 -0.49664640  1.000000000 -0.18897412 -0.39286092
## PETROLEUM        -0.149606971 -0.06454496 -0.188974124  1.00000000 -0.04083260
## HYDRO            -0.276011726 -0.24816204 -0.392860922 -0.04083260  1.00000000
## GEOTHERMAL       -0.173549377 -0.14679276  0.140746497  0.13580890 -0.04107287
## SOLARPV          -0.196679324 -0.28314790  0.244078425  0.09911190 -0.04204189
## WIND             -0.311886710  0.13094496 -0.460727181 -0.06058792  0.16670820
## BIOMASS_OTHER    -0.132359399 -0.37255995  0.066964182  0.06676950  0.18294830
## consumption_Tbtu  0.158536193 -0.10135467  0.219791849 -0.11779548 -0.20542049
##                   GEOTHERMAL     SOLARPV        WIND BIOMASS_OTHER
## NUCLEAR          -0.17354938 -0.19667932 -0.31188671   -0.13235940
## COAL             -0.14679276 -0.28314790  0.13094496   -0.37255995
## NATURAL.GAS       0.14074650  0.24407842 -0.46072718    0.06696418
## PETROLEUM         0.13580890  0.09911190 -0.06058792    0.06676950
## HYDRO            -0.04107287 -0.04204189  0.16670820    0.18294830
## GEOTHERMAL        1.00000000  0.69633847 -0.10781855   -0.05276399
## SOLARPV           0.69633847  1.00000000 -0.20060339    0.32354221
## WIND             -0.10781855 -0.20060339  1.00000000   -0.07437744
## BIOMASS_OTHER    -0.05276399  0.32354221 -0.07437744    1.00000000
## consumption_Tbtu  0.09250805  0.09069217 -0.04690664   -0.18473658
##                  consumption_Tbtu
## NUCLEAR                0.15853619
## COAL                  -0.10135467
## NATURAL.GAS            0.21979185
## PETROLEUM             -0.11779548
## HYDRO                 -0.20542049
## GEOTHERMAL             0.09250805
## SOLARPV                0.09069217
## WIND                  -0.04690664
## BIOMASS_OTHER         -0.18473658
## consumption_Tbtu       1.00000000
```

Interestingly, the highest correlation coefficients come from SOLARPV and GEOTHERMAL (-0.696) as well as WIND and NATURAL.GAS (-0.461). It is not immediately clear why these predictors in particular are correlated. Surprisingly, COAL and NATURAL.GAS have a negative relationship, with a coefficient of -0.497 (see "Results and Discussion".)

We then compute the condition numbers and variance inflation factors (VIFs) for our *stepped* model.

```
#faraway package
vif(GDPfull_step)
```

```
##          NUCLEAR          SOLARPV consumption_Tbtu
##         1.079000         1.061126         1.042048
```

We see that none of the VIFs for the 3 selected predictors from the stepwise regression are large, since they are very close to 1. Thus, we conclude that collinearity isn't a significant issue for our stepped model.

Large leverage points:

```
# hatv1 <- hatvalues(GDPfull_step)
# p <- sum(hatv1)
# n <- 51
# 2*p/n
#
# rev(sort(hatv1))
# fulldata[44,]
# fulldata[5,]
# fulldata[29,]
# fulldata[30,]
```

Texas, California, Nevada, and New Hampshire are significant leverage points because they exceed the critical value of 2p/n = 0.157.

Outliers:

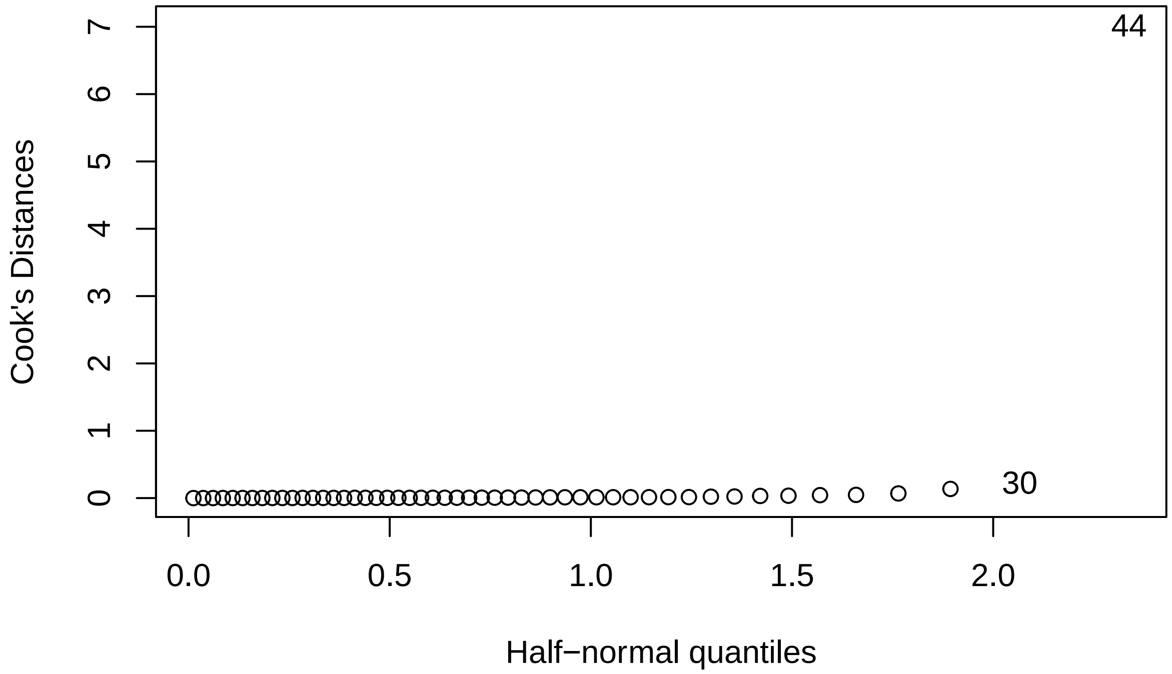
```
stud <- rstudent(GDPfull_step)
stud[which.max(abs(stud))]
```

```
##        44
## -4.760161
```

Texas is an outlier because it exceeds the Bonferroni critical value.

Influential Points:

```
cook <- cooks.distance(GDPfull_step)
halfnorm(cook, ylab="Cook's Distances")
```



```
fulldata[44,] #Texas
```

```
##     X STATE total_GDP NUCLEAR COAL NATURAL.GAS PETROLEUM HYDRO GEOTHERMAL
## 44 44 Texas   1759.73     8.7 16.6        52.7         0   0.4          0
##    SOLARPV WIND BIOMASS_OTHER      party    dem_votes    rep_votes other_votes
## 44     1.7 19.6           0.4 Republican 5,259,126.00 5,890,347.00  165,583.00
##    dem_percent rep_percent other_percent Rank consumption_Tbtu index
## 44        46.5        52.1           1.5   51          14258.8     1
```

```
fulldata[30,] #New Hampshire
```

```
##     X         STATE total_GDP NUCLEAR COAL NATURAL.GAS PETROLEUM HYDRO
## 30 30 New Hampshire     85.11      59  0.8        21.8       0.2   8.9
##    GEOTHERMAL SOLARPV WIND BIOMASS_OTHER    party  dem_votes  rep_votes
## 30          0       0  3.1           6.1 Democrat 424,921.00 365,654.00
##    other_votes dem_percent rep_percent other_percent Rank consumption_Tbtu
## 30   15,607.00        52.7        45.4           1.9    6            324.7
##      index
## 30 0.02277
```

Texas (in particular) and New Hampshire are influential points.

We can try removing Texas–a leverage point, influential point, and outlier–from our model to see how its removal changes the fit of the model.

```
GDPfull_step_Tex <- lm(log(total_GDP) ~ NUCLEAR + SOLARPV + consumption_Tbtu, data = fulldata[-44,])
summary(GDPfull_step_Tex)
```

```
##
## Call:
## lm(formula = log(total_GDP) ~ NUCLEAR + SOLARPV + consumption_Tbtu,
##     data = fulldata[-44, ])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.22803 -0.32523  0.08168  0.33877  1.10564
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      4.212e+00  1.333e-01  31.597  < 2e-16 ***
## NUCLEAR          1.400e-02  4.630e-03   3.024  0.00407 **
## SOLARPV          3.541e-02  2.294e-02   1.544  0.12953
## consumption_Tbtu 5.154e-04  5.589e-05   9.221 5.03e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5349 on 46 degrees of freedom
## Multiple R-squared:  0.7499, Adjusted R-squared:  0.7336
## F-statistic: 45.98 on 3 and 46 DF,  p-value: 6.866e-14
```

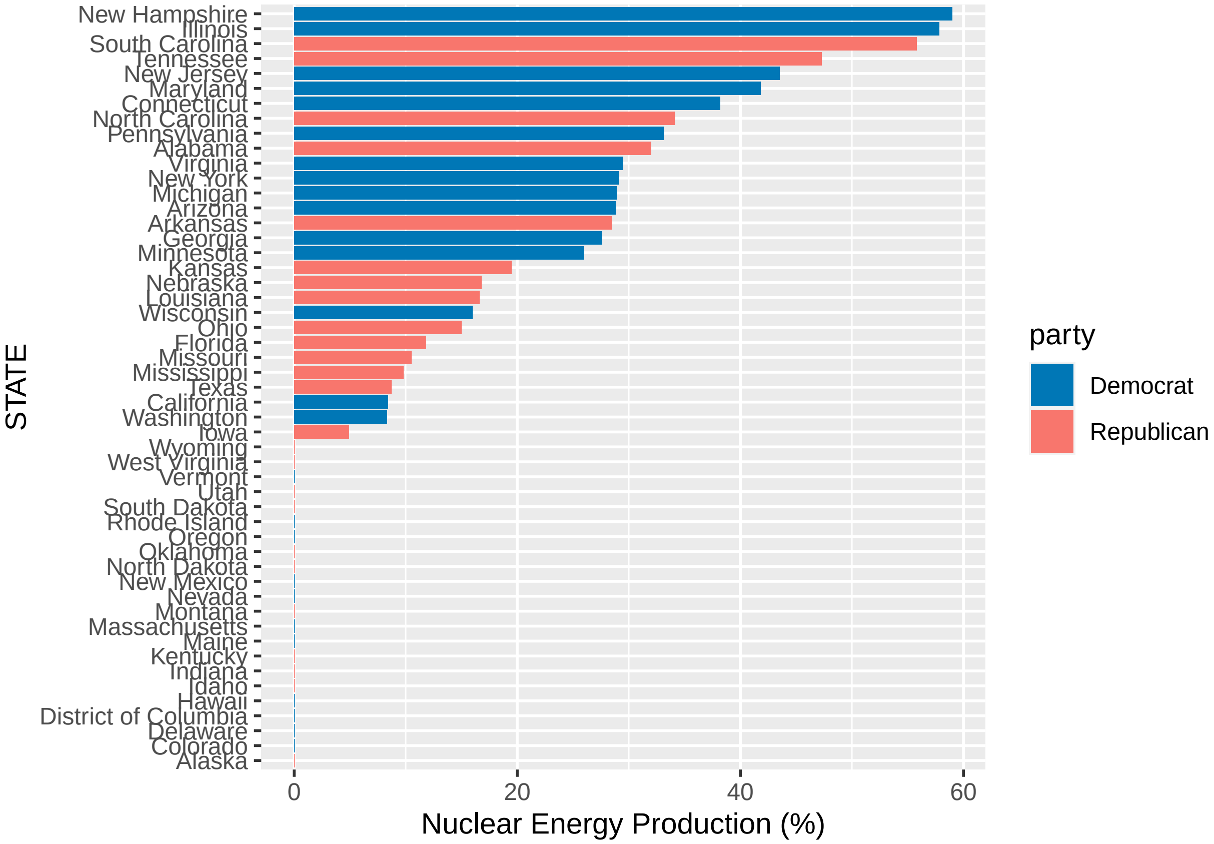
```
summary(GDPfull_step)
```

```
##
## Call:
## lm(formula = log(total_GDP) ~ NUCLEAR + SOLARPV + consumption_Tbtu,
##     data = fulldata)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.51870 -0.33124  0.01246  0.38456  1.24024
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      4.382e+00  1.552e-01  28.245  < 2e-16 ***
## NUCLEAR          2.103e-02  5.304e-03   3.965 0.000248 ***
## SOLARPV          5.995e-02  2.702e-02   2.219 0.031360 *
## consumption_Tbtu 3.033e-04  4.078e-05   7.437 1.78e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6466 on 47 degrees of freedom
## Multiple R-squared:  0.6542, Adjusted R-squared:  0.6321
## F-statistic: 29.64 on 3 and 47 DF,  p-value: 6.609e-11
```

Removing Texas greatly increases the R^2 value, from 0.63 to 0.73. Interestingly, consumption_Tbtu becomes much more significant (its p-value is reduced by almost a factor of 3), but *energy production* predictors become *less* significant. Coefficients for all three predictors change substantially. Residual standard error decreases substantially as well.

Analysis on Outliers, Leverage Points, Influential Points

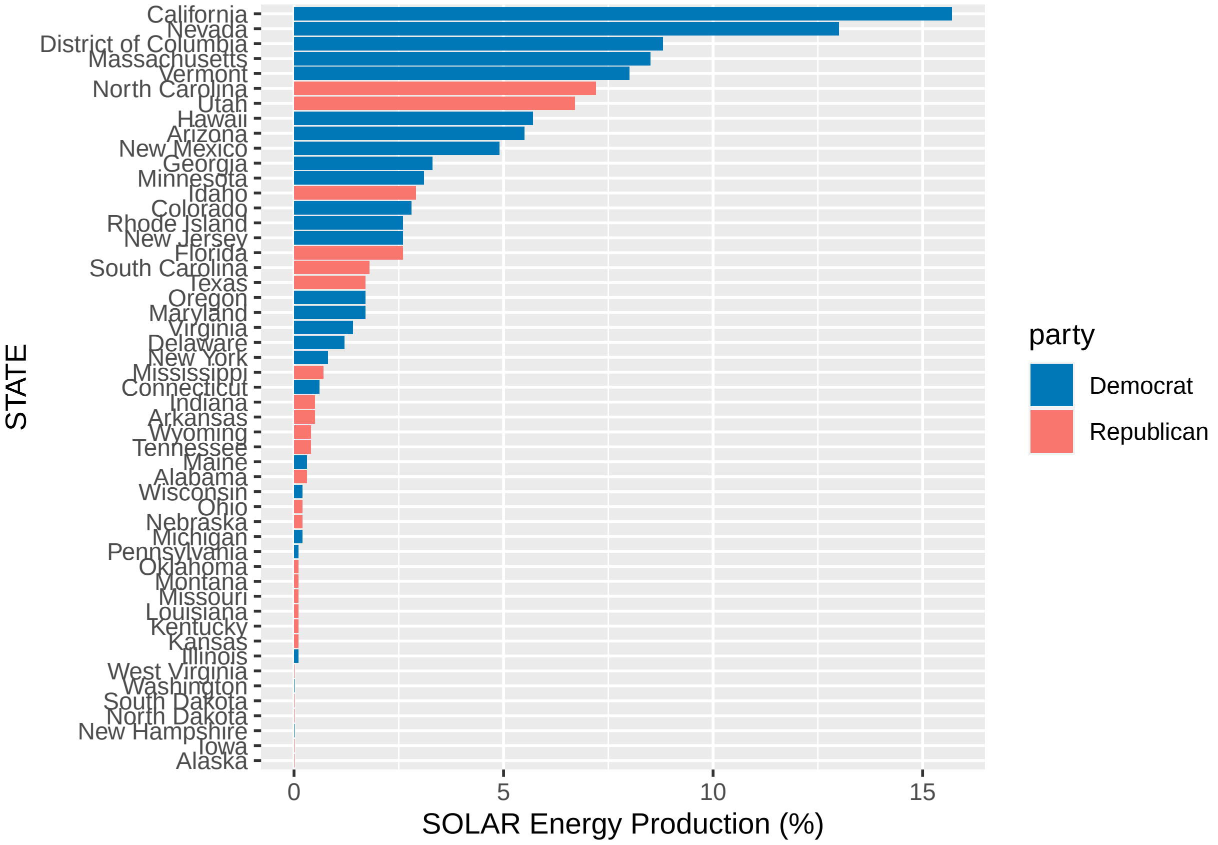
```
#plotting NUCLEAR
ggplot(data = fulldata, mapping = aes(x = reorder(STATE, NUCLEAR), y=NUCLEAR, fill=party)) + scale_fill_
  geom_bar(stat = "identity") + coord_flip()+
  labs(x="STATE", y="Nuclear Energy Production (%)")
```



```
  ggtitle("NUCLEAR")
```

```
## $title
## [1] "NUCLEAR"
##
## attr(,"class")
## [1] "labels"
```

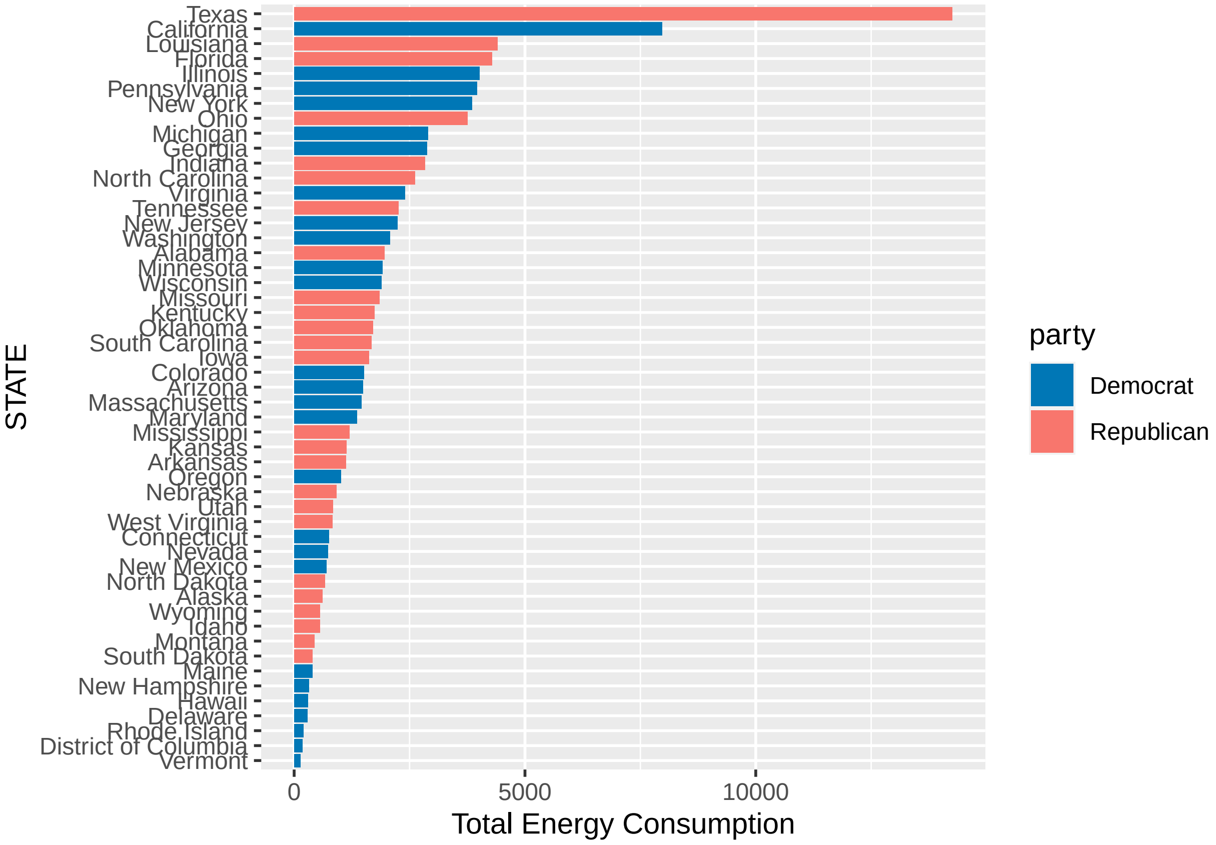
```
#plotting SOLAR
ggplot(data = fulldata, mapping = aes(x = reorder(STATE, SOLARPV), y=SOLARPV, fill=party)) + scale_fill_
  geom_bar(stat = "identity") + coord_flip()+
  labs(x="STATE", y="SOLAR Energy Production (%)")
```



```
  ggtitle("SOLAR")
```

```
## $title
## [1] "SOLAR"
##
## attr(,"class")
## [1] "labels"
```

```
#plotting Consumption
ggplot(data = fulldata, mapping = aes(x = reorder(STATE, consumption_Tbtu), y=consumption_Tbtu, fill=pa
  geom_bar(stat = "identity") + coord_flip()+
  labs(x="STATE", y="Total Energy Consumption")
```



```
  ggtitle("Total Energy Consumption ($)")
```

```
## $title
## [1] "Total Energy Consumption ($)"
##
## attr(,"class")
## [1] "labels"
```

Plotting the graphs for the significant variables, we see that the outliers, influential points, and leverage points calculated above are visible in the plots of GDP against significant predictors. Nuclear energy and solar energy are statistically significant predictors of total GDP. However, as seen from the graphs above, there is a high variability within these variables. About a half of the states have not implemented these as sources of electricity productions, which brings concerns about how accurately our model can predict the total GDP based on only these two predictors. However, we are well aware that there are many other factors that contributes to the total GDP. Our model may not be the best model to predict the total GDP, but it still demonstrates an important trend that shifting towards nuclear and solar energy production type might bring a positive increment in total GDP.

##Mitigating problems with the error (non-constant variance)

```
#insert residuals --fitted plot for full model
```

We chose to refit the full model using the Huber method, in order to down-weight extreme observations without being forced to remove them. We want to be able to keep all 51 state observations in the model,since n = 51 is a relatively small number of observations and since we intend to make conclusions about United States states as a whole.

```
rlmGDPfull <- rlm(log(total_GDP)~ NUCLEAR + COAL+ NATURAL.GAS + PETROLEUM + HYDRO + GEOTHERMAL + SOLARP
summary(rlmGDPfull)
```

```
##
## Call: rlm(formula = log(total_GDP) ~ NUCLEAR + COAL + NATURAL.GAS +
##     PETROLEUM + HYDRO + GEOTHERMAL + SOLARPV + WIND + BIOMASS_OTHER +
##     consumption_Tbtu, data = fulldata)
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.51544 -0.39277  0.07919  0.31163  1.40722
##
## Coefficients:
##                  Value    Std. Error t value
## (Intercept)        1.0906 131.5877     0.0083
## NUCLEAR            0.0488   1.3161     0.0371
## COAL               0.0321   1.3152     0.0244
## NATURAL.GAS        0.0358   1.3156     0.0272
## PETROLEUM          0.0282   1.3147     0.0214
## HYDRO              0.0291   1.3157     0.0221
## GEOTHERMAL        -0.0439   1.3277    -0.0331
## SOLARPV            0.1134   1.3261     0.0855
## WIND               0.0313   1.3168     0.0237
## BIOMASS_OTHER      0.0148   1.3157     0.0113
## consumption_Tbtu   0.0004   0.0000     8.2284
##
## Residual standard error: 0.5117 on 40 degrees of freedom
```

```
#summary(GDPfull)
wts <- rlmGDPfull$w
names(wts) <- row.names(fulldata)
head(sort(wts),12) #weighting Texas a LOT less (0.3); WA less, NY/Vermont 0.6ish
```

```
##        44        48        33        46        51        19        30        38
## 0.2736061 0.4890436 0.6054538 0.6779085 0.7341088 0.7480865 0.7595321 0.8444495
##         6        40        22         1
## 0.8629343 0.8832394 0.9125522 1.0000000
```

Unsurprisingly, the Huber regression method assigns Texas the lowest weight out of all observations.

Prediction

We now refit the stepped model by removing one state at a time to see how well the model predicts GDP for the state that has been removed. First, we predict

```
# Alabama as a test state
GDPreduc_Al <-lm(log(total_GDP) ~ NUCLEAR + SOLARPV + consumption_Tbtu, fulldata[-1,]) #take out WA #ta

testAlabama <-fulldata[1,] %>%
  dplyr::select(NUCLEAR, SOLARPV, consumption_Tbtu)

exp(predict(GDPreduc_Al,testAlabama, interval = 'prediction'))
```

```
##        fit      lwr      upr
## 1 291.9911 76.41415 1115.746
```

```
fulldata[1,]$total_GDP
```

```
## [1] 224.87
```

```r
# Washington as a test state
GDPreduc_WA <-lm(log(total_GDP) ~ NUCLEAR + SOLARPV + consumption_Tbtu, fulldata[-48,]) #take out WA--m

testWash <-fulldata[48,] %>%
  dplyr::select(NUCLEAR, SOLARPV, consumption_Tbtu)

exp(predict(GDPreduc_WA,testWash, interval = 'prediction'))
```

```
##         fit      lwr      upr
## 48 171.0131 47.35903 617.5267
```

```r
fulldata[48,]$total_GDP
```

```
## [1] 618.7
```

Out of curiosity, we also tested whether the stepped model with the Huber method applied was better at predicting observations than the non-Huber model.

```r
#testing whether Huber regression is better at predicting:
rlmGDPreduc_Al <-rlm(log(total_GDP) ~ NUCLEAR + SOLARPV + consumption_Tbtu, fulldata[-1,]) #take out Al
summary(rlmGDPreduc_Al)
```

```
## 
## Call: rlm(formula = log(total_GDP) ~ NUCLEAR + SOLARPV + consumption_Tbtu, 
##     data = fulldata[-1, ])
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -2.68340 -0.34143  0.02822  0.31847  1.15334 
## 
## Coefficients:
##                  Value   Std. Error t value
## (Intercept)       4.2967  0.1395    30.7904
## NUCLEAR           0.0193  0.0048     4.0272
## SOLARPV           0.0505  0.0243     2.0777
## consumption_Tbtu  0.0004  0.0000    10.7317
## 
## Residual standard error: 0.4976 on 46 degrees of freedom
```

The question that drove our linear modeling process was whether looking at a specific sector of economy can provides us some insights into the total GDP. From our linear model, we were able to conclude that there is a positive correlation between the total GDP produced per state and total energy consumption, nuclear energy, and solar energy production types. When we tried to predict a state's GDP with our linear model, however, we realized our model does not perform the best with predctions.

What we were able to conclude, on the other hand, was the trend that shifting towards nuclear and solar energy types than other energy production types–since a percent increase in these variables mean a percent decrease in other production types–may bring more statistically significant, positive impact on the total GDP. However, we still have to take into considerations that we are only looking at the impact of an energy sector to the total GDP from a single year, and that not all states implement nuclear and solar energy as their main electricity production source.

Full Model Extension: Political Affiliation ~ Energy production source by type We also wanted to look at energy production by type (i.e. Solar, wind, petroleum) as predictors for political affiliation. In this case, we can construct models with our breakdown of energy types by state as predictors and the percentage of voters who voted Republican and who voted Democrat.

Below is a model with percentage of democratic voters as a response and energy types as predictors.

```
demMdl<-lm(dem_percent~NUCLEAR+COAL+NATURAL.GAS+PETROLEUM+HYDRO+GEOTHERMAL+SOLARPV+WIND+BIOMASS_OTHER,da
summary(demMdl)
```

```
##
## Call:
## lm(formula = dem_percent ~ NUCLEAR + COAL + NATURAL.GAS + PETROLEUM +
##     HYDRO + GEOTHERMAL + SOLARPV + WIND + BIOMASS_OTHER, data = fulldata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -14.901  -5.687   0.535   4.967  15.304
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     432.847   1612.918   0.268    0.790
## NUCLEAR          -3.814     16.129  -0.236    0.814
## COAL             -4.037     16.120  -0.250    0.804
## NATURAL.GAS      -3.841     16.124  -0.238    0.813
## PETROLEUM        -3.724     16.114  -0.231    0.818
## HYDRO            -3.932     16.127  -0.244    0.809
## GEOTHERMAL       -5.449     16.279  -0.335    0.740
## SOLARPV          -2.445     16.243  -0.151    0.881
## WIND             -3.893     16.137  -0.241    0.811
## BIOMASS_OTHER    -3.104     16.135  -0.192    0.848
##
## Residual standard error: 8.185 on 41 degrees of freedom
## Multiple R-squared:  0.6208, Adjusted R-squared:  0.5376
## F-statistic: 7.458 on 9 and 41 DF,  p-value: 2.335e-06
```

Below is the same model fitted with Republican voter percentage as the response:

```
repMdl<-lm(rep_percent~NUCLEAR+COAL+NATURAL.GAS+PETROLEUM+HYDRO+GEOTHERMAL+SOLARPV+WIND+BIOMASS_OTHER,da
summary(repMdl)
```

```
##
## Call:
## lm(formula = rep_percent ~ NUCLEAR + COAL + NATURAL.GAS + PETROLEUM +
##     HYDRO + GEOTHERMAL + SOLARPV + WIND + BIOMASS_OTHER, data = fulldata)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.4272  -5.0398  -0.5444   5.7467  15.0107
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -332.562   1614.912  -0.206    0.838
## NUCLEAR           3.802     16.149   0.235    0.815
## COAL              4.008     16.140   0.248    0.805
## NATURAL.GAS       3.823     16.144   0.237    0.814
## PETROLEUM         3.699     16.134   0.229    0.820
## HYDRO             3.892     16.147   0.241    0.811
## GEOTHERMAL        5.445     16.299   0.334    0.740
## SOLARPV           2.382     16.263   0.146    0.884
## WIND              3.867     16.157   0.239    0.812
## BIOMASS_OTHER     3.072     16.155   0.190    0.850
```

```
##
## Residual standard error: 8.195 on 41 degrees of freedom
## Multiple R-squared:  0.6169, Adjusted R-squared:  0.5329
## F-statistic: 7.337 on 9 and 41 DF,  p-value: 2.817e-06
```

Since we are using the percentage of votes as our response, our models for the Democrat vs. Republican percentages should yield very similar results, in terms of significant predictors. For now, let's focus on Republican response data. Before conducting additional analysis, let us consider the addition of other economic variables, specifically Total GDP and Total Energy Consumption.

```r
#updating the model
demMdl2<-update(demMdl,.~.+total_GDP +consumption_Tbtu)
repMdl2<-update(repMdl,.~.+total_GDP +consumption_Tbtu)

#summarize both updated models
summary(demMdl2)
```

```
##
## Call:
## lm(formula = dem_percent ~ NUCLEAR + COAL + NATURAL.GAS + PETROLEUM +
##     HYDRO + GEOTHERMAL + SOLARPV + WIND + BIOMASS_OTHER + total_GDP +
##     consumption_Tbtu, data = fulldata)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.3698  -6.0324   0.9047   5.0807  14.0516
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -2.469e+02  1.610e+03  -0.153   0.8790
## NUCLEAR           2.955e+00  1.610e+01   0.183   0.8554
## COAL              2.773e+00  1.610e+01   0.172   0.8641
## NATURAL.GAS       2.966e+00  1.610e+01   0.184   0.8548
## PETROLEUM         3.095e+00  1.609e+01   0.192   0.8485
## HYDRO             2.844e+00  1.610e+01   0.177   0.8607
## GEOTHERMAL        1.427e+00  1.625e+01   0.088   0.9305
## SOLARPV           3.884e+00  1.619e+01   0.240   0.8116
## WIND              2.924e+00  1.612e+01   0.181   0.8570
## BIOMASS_OTHER     3.780e+00  1.611e+01   0.235   0.8157
## total_GDP         9.540e-03  4.343e-03   2.197   0.0341 *
## consumption_Tbtu -1.658e-03  9.405e-04  -1.763   0.0858 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.916 on 39 degrees of freedom
## Multiple R-squared:  0.6626, Adjusted R-squared:  0.5674
## F-statistic: 6.963 on 11 and 39 DF,  p-value: 2.427e-06
```

```r
summary(repMdl2)
```

```
##
## Call:
## lm(formula = rep_percent ~ NUCLEAR + COAL + NATURAL.GAS + PETROLEUM +
##     HYDRO + GEOTHERMAL + SOLARPV + WIND + BIOMASS_OTHER + total_GDP +
##     consumption_Tbtu, data = fulldata)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.7664  -5.1196  -0.5082   5.8464  13.1507
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        3.701e+02  1.616e+03   0.229   0.8201
## NUCLEAR           -3.200e+00  1.616e+01  -0.198   0.8441
## COAL              -3.032e+00  1.616e+01  -0.188   0.8521
## NATURAL.GAS       -3.215e+00  1.616e+01  -0.199   0.8433
## PETROLEUM         -3.347e+00  1.615e+01  -0.207   0.8369
## HYDRO             -3.114e+00  1.616e+01  -0.193   0.8482
## GEOTHERMAL        -1.658e+00  1.631e+01  -0.102   0.9196
## SOLARPV           -4.198e+00  1.624e+01  -0.258   0.7974
## WIND              -3.182e+00  1.617e+01  -0.197   0.8451
## BIOMASS_OTHER     -4.032e+00  1.617e+01  -0.249   0.8044
## total_GDP         -9.370e-03  4.359e-03  -2.150   0.0378 *
## consumption_Tbtu   1.705e-03  9.439e-04   1.807   0.0785 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.944 on 39 degrees of freedom
## Multiple R-squared:  0.6576, Adjusted R-squared:  0.561
## F-statistic: 6.808 on 11 and 39 DF,  p-value: 3.141e-06
```

We see that total_GDP is a significant predictor and our R-squared value has increased, as might be expected with the addition of new predictors. However, we still find that most of our predictors are not very significant. To determine if we should, in fact, include these new predictors in the updated model, we can run a analysis of variance on both the models.

```
# republican
anova(repMdl,repMdl2)
```

```
## Analysis of Variance Table
##
## Model 1: rep_percent ~ NUCLEAR + COAL + NATURAL.GAS + PETROLEUM + HYDRO +
##     GEOTHERMAL + SOLARPV + WIND + BIOMASS_OTHER
## Model 2: rep_percent ~ NUCLEAR + COAL + NATURAL.GAS + PETROLEUM + HYDRO +
##     GEOTHERMAL + SOLARPV + WIND + BIOMASS_OTHER + total_GDP +
##     consumption_Tbtu
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     41 2753.2
## 2     39 2461.3  2    291.94 2.3129 0.1124
```

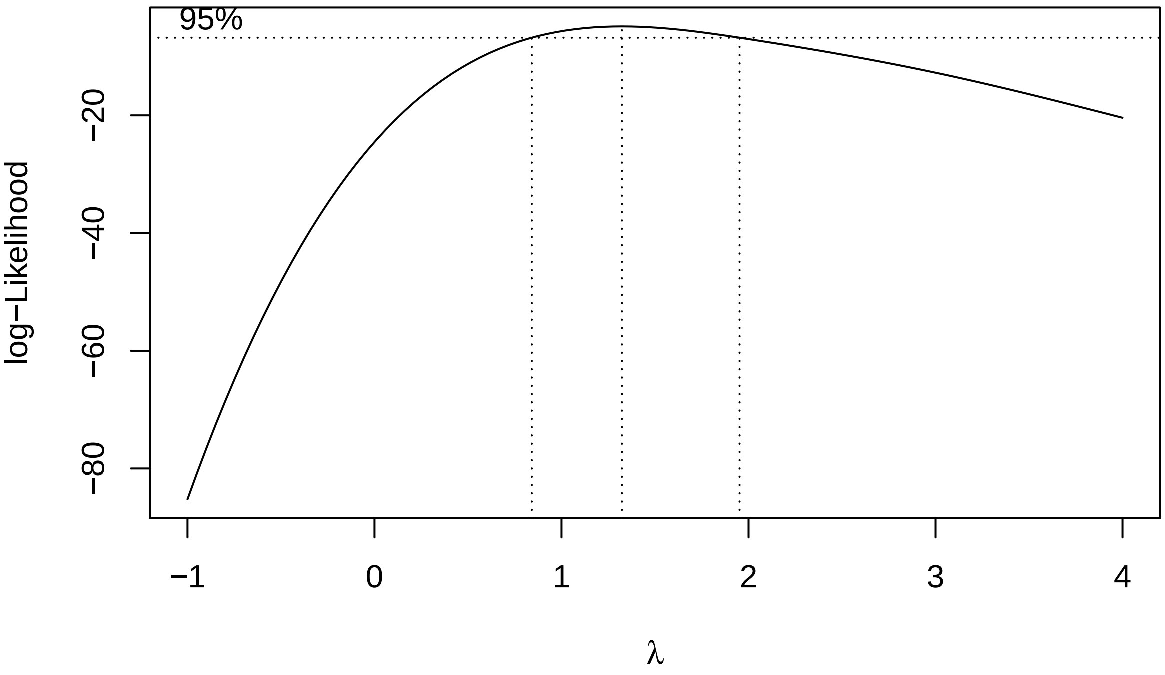
```
# democrat
anova(demMdl,demMdl2)
```

```
## Analysis of Variance Table
##
## Model 1: dem_percent ~ NUCLEAR + COAL + NATURAL.GAS + PETROLEUM + HYDRO +
##     GEOTHERMAL + SOLARPV + WIND + BIOMASS_OTHER
## Model 2: dem_percent ~ NUCLEAR + COAL + NATURAL.GAS + PETROLEUM + HYDRO +
##     GEOTHERMAL + SOLARPV + WIND + BIOMASS_OTHER + total_GDP +
##     consumption_Tbtu
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     41 2746.4
## 2     39 2443.7  2    302.71 2.4155 0.1026
```

Our null hypothesis for the ANOVA comparison is that there is not significant difference in the predictive power of the small model and the large model that includes the economic predictors. Our p-value for both comparisons is greater than 0.1, disallowing us from rejecting the null hypothesis. Therefore, we can use our smaller models repMdl', which include only energy production by type as predictors. This also focuses our analysis on energy type.

##Transformations We want to determine if a transformation of our response data is needed. We can use a Box-Cox analysis:

```
boxcox(repMdl,plotit=T,lambda = seq(-1,4))
```



Our peak is centered around 1.3, although 1 is still contained in our confidence interval, suggesting a transformation is not likely necessary. Thus, we do not apply a transformation of the response.

##Stepwise regression It is unclear if all of our predictors are necessary to construct a useful model. We use a stepwise regression to evaluate which predictors are most impactful. The stepwise regression tells us that SOLAR and BIOMASS can be removed from the model, which is a surprising deviation from our first model with GDP as the response. We construct an updated version of repMdl with these predictors removed:

```
step(repMdl)
```

```
## Start:  AIC=223.42
## rep_percent ~ NUCLEAR + COAL + NATURAL.GAS + PETROLEUM + HYDRO + 
##     GEOTHERMAL + SOLARPV + WIND + BIOMASS_OTHER
## 
##                 Df Sum of Sq    RSS    AIC
## - SOLARPV        1    1.4411 2754.7 221.45
## - BIOMASS_OTHER  1    2.4290 2755.7 221.47
## - PETROLEUM      1    3.5305 2756.8 221.49
## - NUCLEAR        1    3.7212 2757.0 221.49
## - NATURAL.GAS    1    3.7655 2757.0 221.49
## - WIND           1    3.8460 2757.1 221.50
## - HYDRO          1    3.9016 2757.2 221.50
## - COAL           1    4.1404 2757.4 221.50
## - GEOTHERMAL     1    7.4938 2760.7 221.56
## <none>                        2753.2 223.42
##
## Step:  AIC=221.45
## rep_percent ~ NUCLEAR + COAL + NATURAL.GAS + PETROLEUM + HYDRO +
##     GEOTHERMAL + WIND + BIOMASS_OTHER
##
##                 Df Sum of Sq    RSS    AIC
## - BIOMASS_OTHER  1     66.49 2821.2 220.67
## <none>                       2754.7 221.45
## - GEOTHERMAL     1    271.79 3026.5 224.25
## - PETROLEUM      1    385.24 3139.9 226.13
## - NATURAL.GAS    1    463.91 3218.6 227.39
## - NUCLEAR        1    468.57 3223.3 227.46
## - WIND           1    505.77 3260.5 228.05
## - HYDRO          1    525.59 3280.3 228.36
## - COAL           1    605.21 3359.9 229.58
##
## Step:  AIC=220.67
## rep_percent ~ NUCLEAR + COAL + NATURAL.GAS + PETROLEUM + HYDRO +
##     GEOTHERMAL + WIND
##
##               Df Sum of Sq    RSS    AIC
## <none>                     2821.2 220.67
## - GEOTHERMAL   1    327.59 3148.8 224.27
## - PETROLEUM    1    888.82 3710.0 232.63
## - NATURAL.GAS  1   1598.53 4419.7 241.56
## - WIND         1   1669.24 4490.4 242.37
## - HYDRO        1   1730.93 4552.1 243.07
## - NUCLEAR      1   1735.31 4556.5 243.12
## - COAL         1   2820.58 5641.8 254.01
```

```
##
## Call:
## lm(formula = rep_percent ~ NUCLEAR + COAL + NATURAL.GAS + PETROLEUM +
##     HYDRO + GEOTHERMAL + WIND, data = fulldata)
##
## Coefficients:
## (Intercept)      NUCLEAR         COAL  NATURAL.GAS    PETROLEUM        HYDRO
##    -45.0521       0.9271       1.1277       0.9413       0.8302       1.0229
##  GEOTHERMAL         WIND
##      1.7794       0.9943
```

```
stepRepMdl<-lm(formula = rep_percent ~ NUCLEAR + COAL + NATURAL.GAS + PETROLEUM +
    HYDRO + GEOTHERMAL + WIND, data = fulldata)
summary(stepRepMdl)
```

```
##
## Call:
## lm(formula = rep_percent ~ NUCLEAR + COAL + NATURAL.GAS + PETROLEUM +
##     HYDRO + GEOTHERMAL + WIND, data = fulldata)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.4592  -4.7270  -0.2815   5.3405  14.3041
```

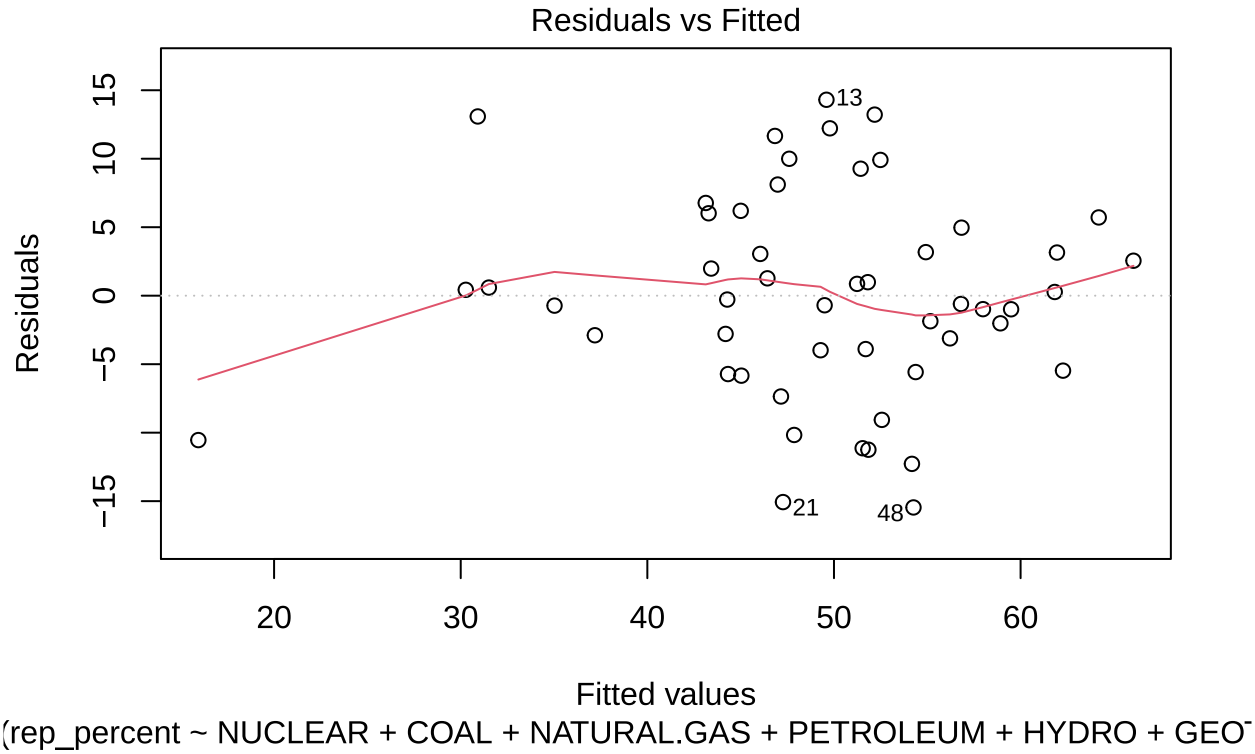
```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -45.0521    17.1007  -2.635 0.011666 *
## NUCLEAR       0.9271     0.1803   5.143 6.33e-06 ***
## COAL          1.1277     0.1720   6.557 5.66e-08 ***
## NATURAL.GAS   0.9413     0.1907   4.936 1.25e-05 ***
## PETROLEUM     0.8302     0.2256   3.681 0.000644 ***
## HYDRO         1.0229     0.1991   5.136 6.47e-06 ***
## GEOTHERMAL    1.7794     0.7963   2.235 0.030698 *
## WIND          0.9943     0.1971   5.044 8.76e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.1 on 43 degrees of freedom
## Multiple R-squared:  0.6075, Adjusted R-squared:  0.5436
## F-statistic: 9.507 on 7 and 43 DF,  p-value: 4.487e-07
```

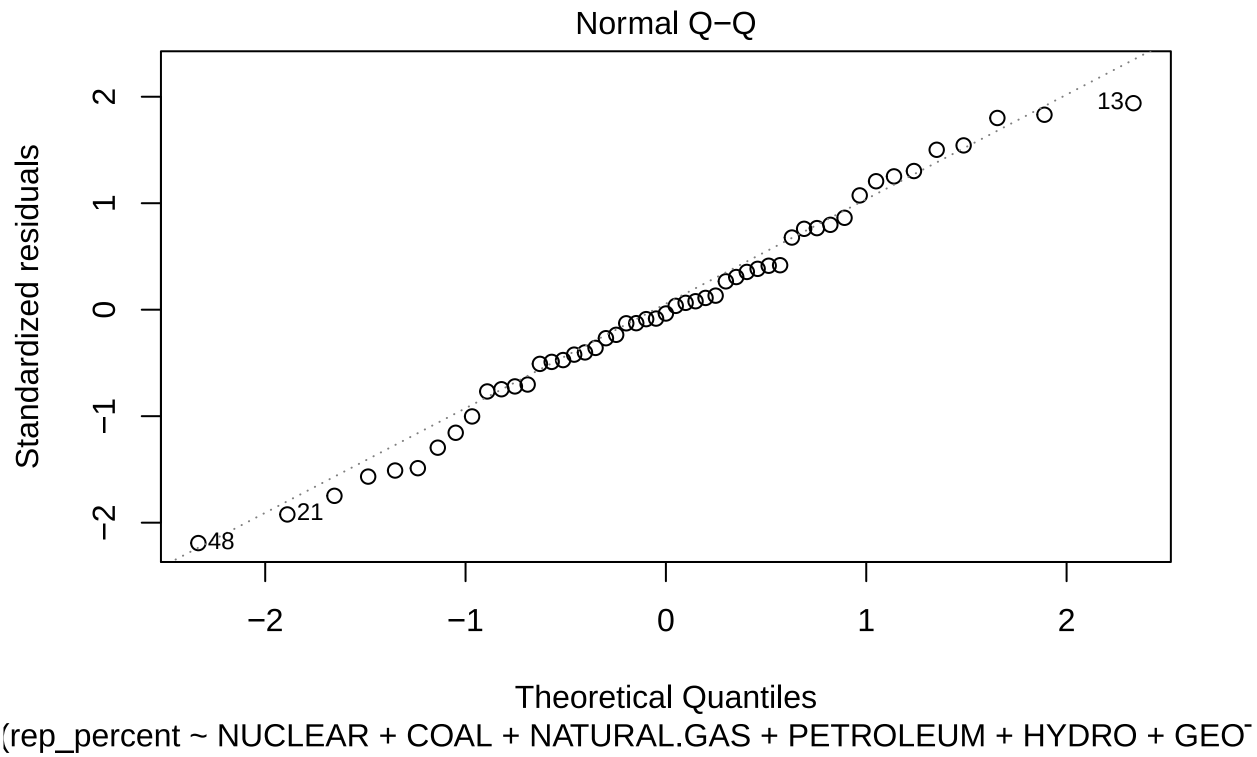
Removing SOLARPV and BIOMASS_OTHER shows that our R-squared value has increased, and we now see all of our predictors are significant at the 5% level or better.

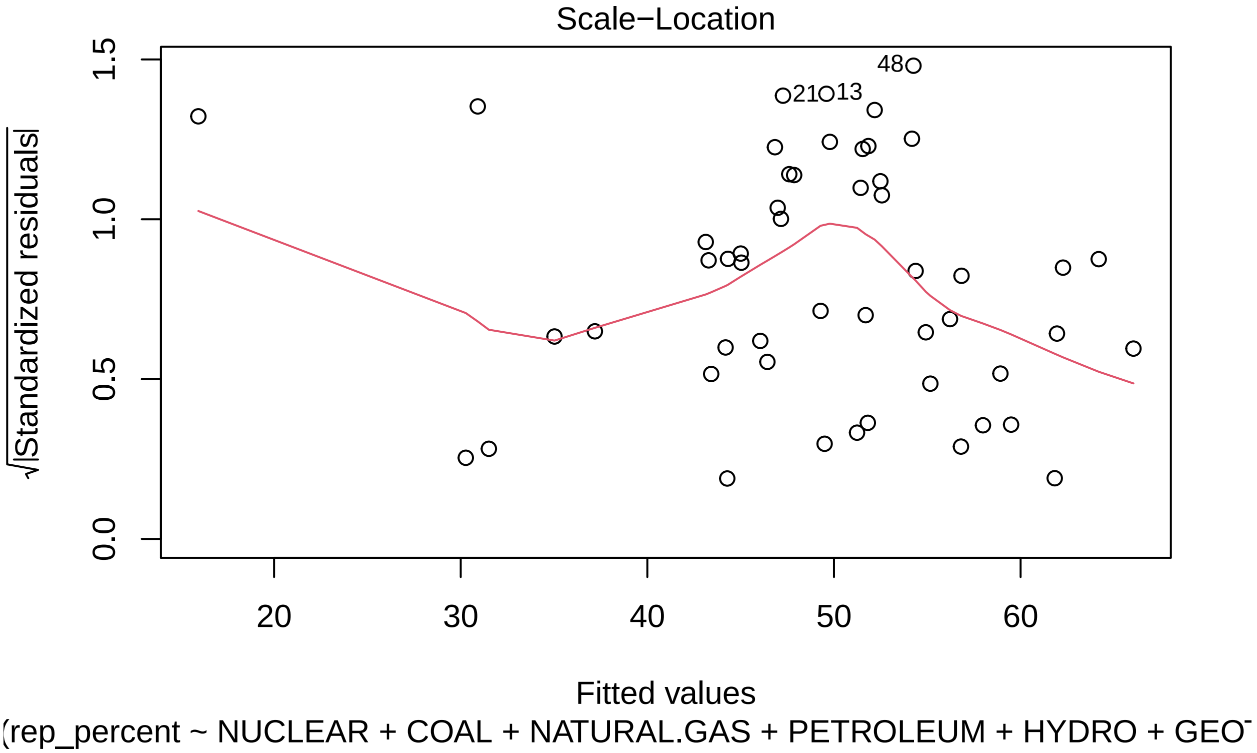
Model Diagnostics

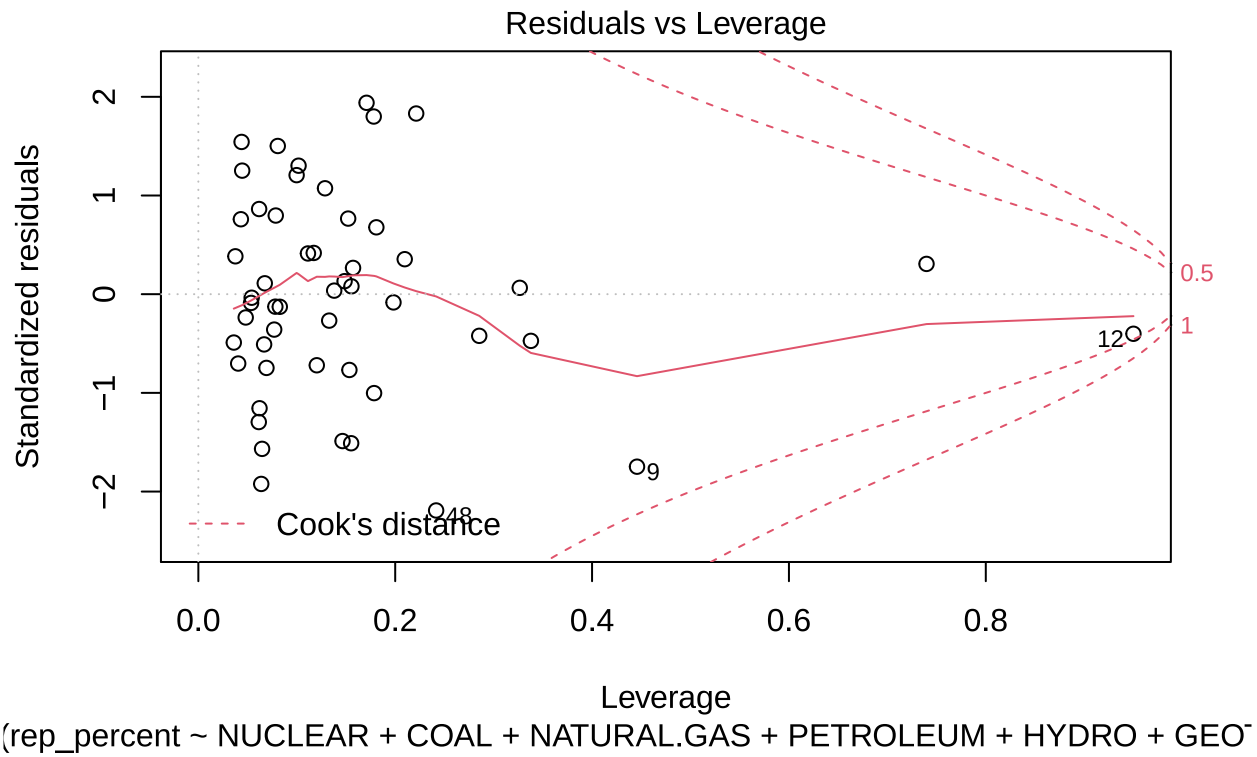
We can use the plot function to look at our residuals and check for leverage points. We can conduct further analysis if necessary.

```
plot(stepRepMdl)
```




Normal Q−Q
Standardized residuals
2
1
0
−1
−2
13
21
48
−2
−1
0
1
2
Theoretical Quantiles
(rep_percent ~ NUCLEAR + COAL + NATURAL.GAS + PETROLEUM + HYDRO + GEO


Scale−Location
√|Standardized residuals|
1.5
1.0
0.5
0.0
48
21
13
20
30
40
50
60
Fitted values
(rep_percent ~ NUCLEAR + COAL + NATURAL.GAS + PETROLEUM + HYDRO + GEO



```
shapiro.test(residuals(stepRepMdl))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(stepRepMdl)
## W = 0.98006, p-value = 0.5414
```

Using the plot function, we see that our residuals seem mostly centered around zero, and we do not see evidence of large leverage points that surpass the cook's distance interval, however we see a few points that are close, so we conduct explicit analysis below. We additionally see from our QQ plot that our residuals largely follow a normal distribution. Our model also passes the Shapiro Wilk test, with a high p-value. Therefore, we fail to reject the null hypothesis that our residuals follow a normal distribution.

Large leverage points.

We can assess for large leverage points explicitly following the same method as above in the GDP model.

```
hatv2 <- hatvalues(stepRepMdl)
p <- sum(hatv2)
n <- 51
2*p/n
```

```
## [1] 0.3137255
```

```
head(rev(sort(hatv2)))
```

```
##        12        29         9        16        46         5
## 0.9500188 0.7397739 0.4456751 0.3378615 0.3265392 0.2853606
```

We see a number of potential leverage points that exceed our threshold value of 0.313. Namely, Hawaii, Nevada, Washington D.C., Iowa and Vermont exceed this threshold. Since the leverage points do not necessarily represent deviations from the trend of the model, we do not remove them here, but we do check for outliers and more broad influential points.

Outliers

```
stud2 <- rstudent(stepRepMdl)
stud2[which.max(abs(stud2))]
```

```
##        48
## -2.298075
```

```
qt(0.05/(n*2),n-p)
```

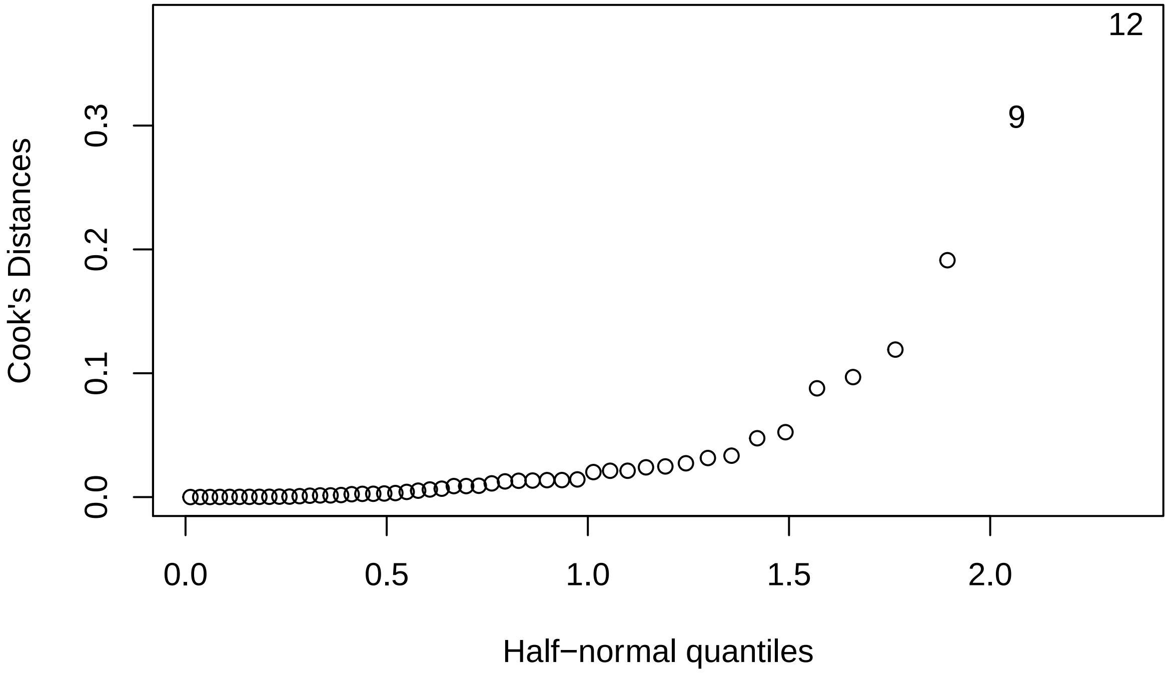
```
## [1] -3.538394
```

```
# rev(sort(abs(stud2)))
```

Based on the Bonferroni critical value, we do not see any significant outliers. Washington has the largest absolute studentized value at 2.298, which does not exceed the Bonferroni critical value of 3.538.

Influential Points

```
cook2 <- cooks.distance(stepRepMdl)
halfnorm(cook2, ylab="Cook's Distances")
```



```
fulldata[9,] #Washington D.C.
```

```
##   X                STATE total_GDP NUCLEAR COAL NATURAL.GAS PETROLEUM HYDRO
## 9 9 District of Columbia    132.53       0    0        64.8         0     0
##   GEOTHERMAL SOLARPV WIND BIOMASS_OTHER    party  dem_votes rep_votes
## 9          0     8.8    0          26.4 Democrat 317,323.00 18,586.00
##   other_votes dem_percent rep_percent other_percent Rank consumption_Tbtu
## 9    8,447.00        92.1         5.4           2.5    2            174.5
##     index
## 9 0.01224
```

```
fulldata[12,] #Hawaii
```

```
##     X  STATE total_GDP NUCLEAR COAL NATURAL.GAS PETROLEUM HYDRO GEOTHERMAL
## 12 12 Hawaii     89.86       0 12.5           0      66.1     1        2.2
##    SOLARPV WIND BIOMASS_OTHER    party  dem_votes  rep_votes other_votes
## 12     5.7  6.2           6.3 Democrat 366,130.00 196,864.00   11,475.00
##    dem_percent rep_percent other_percent Rank consumption_Tbtu   index
## 12        63.7        34.3             2    5            292.9 0.02054
```

We see that based on the calculation of Cook's distance, Washington D.C. and Hawaii are influential points.

##Predicting with our model on one state's data We repeat the prediction analysis as in the above model to determine the predictive power of our model. Here, we remove Indiana from the analysis and predict the percentage of votes that went Republican based on our model. Our predicted value of 58.06% is very close to the actual value of 57, which also is within the 95% confidence interval.

```
n=15
repMdlPredict <-lm(rep_percent~NUCLEAR+COAL+NATURAL.GAS+PETROLEUM+HYDRO+GEOTHERMAL+WIND, fulldata[-n,])
#summary(repMdlPredict)
testStatePolit <-fulldata[n,]
testStatePolit<- testStatePolit%>%
  dplyr::select(NUCLEAR,COAL,NATURAL.GAS,PETROLEUM,HYDRO,GEOTHERMAL,WIND)

predict(repMdlPredict,testStatePolit,interval = "prediction")
```

```
##         fit      lwr      upr
## 15 58.06628 40.84211 75.29046
```

```
fulldata[n,]$rep_percent
```

```
## [1] 57
```

##Limitation: Collinearity We've already looked at correlation coefficients between predictors for the full GDP model, finding that there is a negative relationship between coal and natural gas.

Beyond correlation coefficients for all predictors, we also want to look for evidence of collinearity amongst predictors in the stepped Republican voter model. We can do this by computing VIFs, which will show the square of the factor of increase in variance as a result of collinearity.

```
vif(stepRepMdl)
```

```
##     NUCLEAR        COAL NATURAL.GAS   PETROLEUM       HYDRO  GEOTHERMAL
##    7.942557   12.246769   17.199471    3.465173    9.734575    1.345929
##        WIND
##    4.916270
```

Our VIF calculation indicates evidence of a fairly high amount of collinearity. Coal and natural gas, in particular show evidence of collinearity, as noted in our correlation matrix (see "Full Model 1"). However, removing either one of these variables decreases the overall performance of the model.

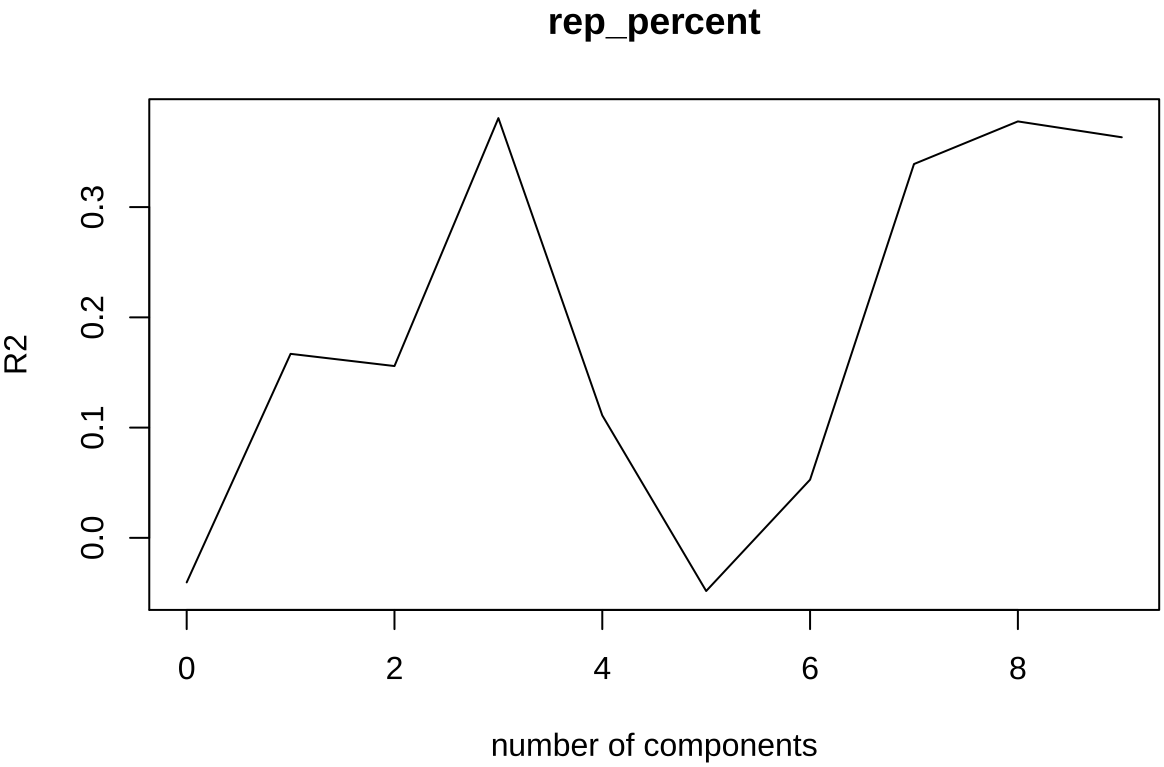
##Principal Component Regression We are also interested in reducing dimensionality and effects of collinearity in our model. We perform a principal components regression (PCR) which makes use of principal component analysis (PCA) to construct a model based on "components" that explain some amount of variability in the data, comprised of linear combinations of our predictors.

```
# Constructing a PCR model:
set.seed (123)
pcr_model <- pcr(rep_percent~NUCLEAR+COAL+NATURAL.GAS+PETROLEUM+HYDRO+GEOTHERMAL+SOLARPV+WIND+BIOMASS_O
summary(pcr_model)
```

```
## Data:    X dimension: 51 9
##  Y dimension: 51 1
## Fit method: svdpc
## Number of components considered: 9
##
## VALIDATION: RMSEP
## Cross-validated using 10 random segments.
##        (Intercept)  1 comps  2 comps  3 comps  4 comps  5 comps  6 comps
## CV           12.11    10.84    10.91    9.342    11.19    12.15    11.55
## adjCV        12.11    10.75    10.82    9.268    10.97    11.89    11.40
##        7 comps  8 comps  9 comps
## CV       9.651    9.365    9.472
## adjCV    9.529    9.250    9.352
##
## TRAINING: % variance explained
##              1 comps  2 comps  3 comps  4 comps  5 comps  6 comps  7 comps
## X              25.55    45.44    61.05    72.43    82.12    90.08    97.76
## rep_percent    37.28    38.71    50.05    52.38    54.31    54.96    59.47
##              8 comps  9 comps
## X             100.00   100.00
## rep_percent    61.64    61.69
```

Plotting our results:

```
validationplot(pcr_model,val.type = "R2")
```



We see that using 5 principal components reduces our root mean squared error, suggesting that this could be a desirable alternative to the OLS model for future work, since this would incorporate multiple predictor variables into a smaller number of linearly-combined components.

IV. RESULTS AND DISCUSSION There was substantial collinearity present in our predictor variables, with some interesting relationships. COAL and NATURAL GAS were negatively correlated, which could make sense if a state is dependent on fracking (a process that extracts natural gas) but also coal production, which in some places are in direct competition with one another. Alternatively, this relationship may occur because coal and natural gas tend to hold higher proportions of electricity consumption than other forms of electricity production. Therefore, an increase or decrease in either may correlate an increase or decrease overall in electricity consumption by state.

While collinearity was mitigated through stepwise regression methods, we cannot be sure that the most useful predictors weren't inadvertently taken out through the stepwise algorithm. We used stepwise regression as an easy way to remove non-significant predictor variables for all of our models, but this was done at the cost of potentially missing important predictors and created smaller models than is potentially ideal. We note that Faraway discourages the use of stepwise and backwards regression except in the case of simple model comparisons or "highly structured heirarchical [sic] models" (Faraway 153), and so while we use ANOVA to determine how much faith we can have in our stepwise models, we cannot say with confidence that our stepwise models are the best variable reduction option. For future research, we recommend utilizing shrinkage methods more fully, such as PCR (which we did incorporate for our political affiliation model) and PLS.

We found that the model with log(GDP) as a response and both electricity energy production and consumption variables as predictors had the best predictive power, compared to either production or consumption variables by themselves. However, since the response is log-transformed, this makes direct interpretation of our model more difficult. We have to exponentiate our coefficients:

```
coef(GDPfull_step)
```

```
##      (Intercept)          NUCLEAR          SOLARPV consumption_Tbtu
##      4.382414870      0.021030524      0.059946789      0.000303274
```

```
exp(coef(GDPfull_step))
```

```
##      (Intercept)          NUCLEAR          SOLARPV consumption_Tbtu
##        80.031065         1.021253         1.061780         1.000303
```

For instance, we see from the above exponentiated coefficients that a 1-percent increase in nuclear production results in a 1.02 billion dollar increase in GDP, while for solar it is a 1.06 billion dollar increase.

For our GDP model, we found that Texas is a problematic observation, since it is a leverage point, influential point, and outlier. This is because the state consumes *by far* the most energy relative to its GDP. After removing Texas from our GDP model post-stepwise regression, the model fit was far better, and the p-values of the predictors changed substantially. However, unlike datasets that involve distinct datapoints that do not interact with each other in a larger system (such as patients, a collection of chemical isomers, or types of tea), we wanted to make conclusions about not only US states in isolate, but also the United States as a whole. Thus, we were hesitant to remove Texas permanently from our dataset, even if we did sacrifice achieving the best fit for our model.

Even after transforming GDP, our residuals were still not random. This is why we implemented the Huber regression method, so that the influence of extreme observations (such as Texas!) non-random errors . Interestingly, even though Washington (observation # 48) does not show up on any of our diagnostic plots, the Huber method significantly downweights this observation. This may be because Washington uses no solar energy for electricity but uses the most hydroelectric power (66%) out of any state, and has unusually high GDP for a state (618B, vs the median of 242B).

For our political affiliation model, were able to conclude that electricity production type can have moderate predictive power for political affiliation in a given state. In particular, we find that predictions of Republican vote percentages are better than expected and are in fact more accurate than our GDP model, to our surprise. It is interesting that total consumption is not a significant predictor, since it implies that even highly Democratic states (such as California) use significant amounts of energy, calling attention to the fact that energy (over)use is a bipartisan issue.

More so than in our GDP model, since more energy production predictors are included, collinearity is more of a problem, and so it is unclear if we can meaningfully interpret the coefficient values of our ordinary least squares model (even though in this case, our response was not log-transformed). Still, as a proof of concept analysis to test the influence of the energy sector on political outcomes, we see that an arguably significant relationship exists between political affiliation and types of energy production.

V. CONCLUSION

A main limitation of this analysis stemmed from a lack of data availability. While finding total GDP breakdowns and 2020 election outcomes by US state was straightforward, we were unable to find energy production data that was not limited merely to electricity energy production by percent shares. This data set also combined energy outputs attributed to biomass *and* other forms of energy production into one variable ("BIOMASS_OTHER"), which made interpretation of the importance of this variable somewhat more difficult, since in fact a few states (such as Maine) actually produce a significant portion of their energy through biomass burning. Furthermore, we were only able to find *total* energy consumption data, and only for the year 2018–all our other data sets were from 2020. Datasets that broke down energy consumption by energy type (renewable/non-renewable etc.) eluded us, even after extensive searching on the Department of Energy website. This issue might be overcome in future research through compiling individual state data sets that may not otherwise be easily available in a 51-state/DC data set.

A more conceptual limitation of our extension model is that political affiliation trends don't vary primarily by state, but rather by population density (in other words, cities are overwhelmingly Democratic, while rural areas are primarily Republican). Thus, future research might include similar predictor/response variables but broken down to the county-level, in order to make more meaningful conclusions about relationships between political outcomes and energy use/production.

In conclusion, nuclear and solar production along with total energy consumption ended up being the most significant variables for predicting log(total GDP) of US states. Predictive power of our model was fairly high, but removing outlier states, especially Texas, improved our model drastically. Some influential observations such as Washington were hidden and only 'discovered' through the Huber method that was implemented as an error mitigation technique. For our political affiliation model, energy production variables ended up being important predictors, while energy consumption was not, and ended up having surprising predictive power, albeit with significant collinearity problems.