Oil Consumption and a Nation's Military

Tyler Nguyen

3/18/2023

Introduction

Research Question In this report, I sought to answer how a nation's military affects its oil consumption. Military emissions are not required to be reported to the United Nations, and I wish to explore if the size of a nation's military is associated with a nation's oil consumption.

Background and Source of Data Oil is the most desired natural resource in the world because of its significance in all areas of our lives, from the gas in our cars, to the heating of our homes, to the production of electricity. I chose to conduct a report using metrics of a nation's military and their subsequent oil consumption. After reading an article which stated that the single largest use of fossil fuels in the world was via the US military, I thought of using my recently acquired multiple linear regression knowledge to further analyze this claim. I pulled data from Kaggle of 'Military Power by Country 2022' which pulled its data from Global Fire Power's index of militaries, which ranked military strength based off of 40 metrics. Oil consumption was one of these metrics, and I chose seven others to serve as the explanatory variables to create a multiple linear regression model which predicted oil consumption.

Methodology and Paper Overview To start, I will conduct an introductory analysis of the variables before building the models in order to observe potential relationships. Afterwards, I will then fit a multiple linear regression model using all variables in the data set, and from this initial model, I will observe the four diagnostic plots to observe if any conditions are violated. Namely, these conditions include linearity, constant variance, and normality of errors. If they are violated, I will transform and refit the data until these assumptions are met. If there is an issue with multicollinearity, I will perform variable selection to reduce the model. Finally, I will interpret the results and analyze which model performs best using my intuition to see if the model is applicable to the real world.

Data Description The dataset is of size [140,8] and contains the following variables:

- Oil.Consumption : A nation's oil consumption in BBL (barrel of crude oil). 1 BBL = 42 gallons
- power_index : A nation's military score based upon factors. Lower score means a nation has a stronger military, and a score of 0 is the theoretic "perfect" military
- Active.personnel : The number of people in nation's military full time
- Armored.Vehicles : The number of armored vehicles of a nation
- Helicopters : The number of total military helicopters of a nation
- Defense.Budget : The defense budget of a nation in Billions of USD
- Navy.ships : The number of warships and submarines of a nation
- Total.Aircraft.Strength: The total number of aircrafts (fighter jets, trainers, transports, bombers, etc) of a nation

There are no NA values in the data frame.

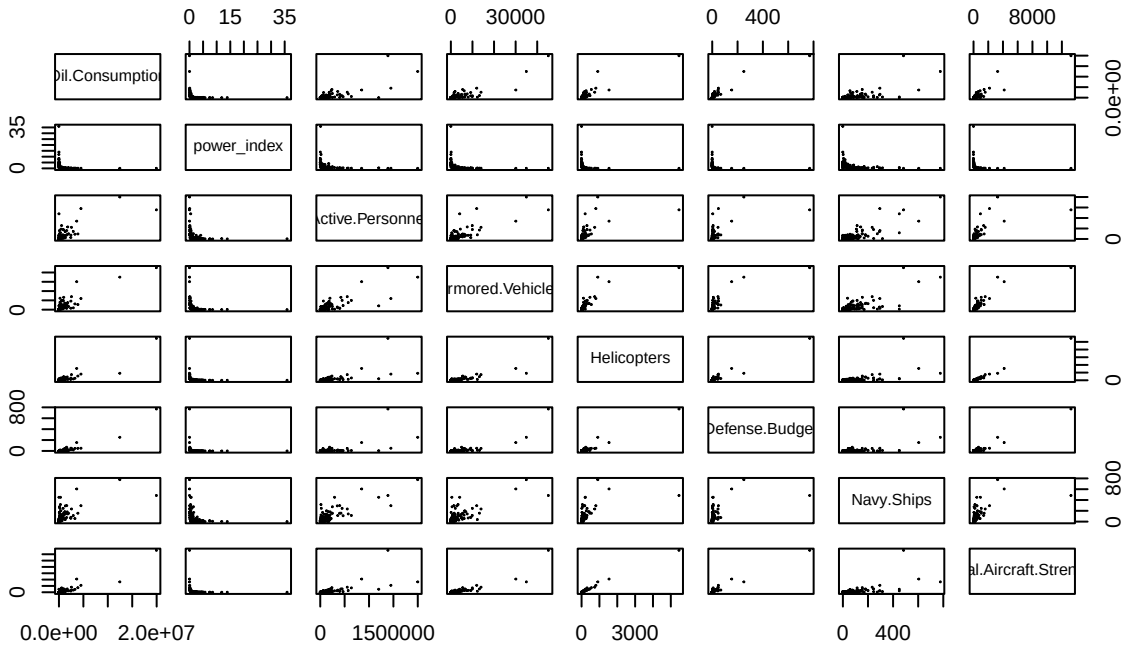
Summary Stats Here are the min, mean, max, and standard deviations of each explanatory variable:

```
##      power_index Active.Personnel Armored.Vehicles Helicopters Defense.Budget
## min         0.05              0.0            20.00        0.00           0.01
## mean        2.14         137857.9          2770.43      141.44          14.47
## max        35.90        2000000.0         45193.00     5463.00         770.00
## sd          3.55         279003.5          5907.14      491.54          69.68
##      Navy.Ships Total.Aircraft.Strength
## min        0.00                   0.00
## mean      75.28                 378.76
## max      777.00               13247.00
## sd       119.58                1222.93
```

Correlation Coefficients Here are the correlation coefficients between each variable and oil consumption

```
##         Oil.Consumption             power_index        Active.Personnel
##               1.0000000              -0.1676239               0.7487992
##        Armored.Vehicles             Helicopters          Defense.Budget
##               0.8685267               0.8972694               0.9399200
##             Navy.Ships Total.Aircraft.Strength
##               0.6342606               0.9204212
```

Scatterplot Matrix The Scatterplot Matrix:



It is seen from the matrix that the relationship between oil consumption and power_index appears to be negatively associated. In fact, it looks very logarithmic, indicating that a potential transformation is

necessary. The variables helicopters, defense budget, and total aircraft strength appear to be positively associated, while Active Personnel, armored vehicles, and navy ships, appear to have a very slight positive correlation.

Models

First Model: Untransformed Full Multiple Linear Regression My first model is a multiple linear regression model including all seven variables with no transformations. This is the equation:

```
##
## Call:
## lm(formula = Oil.Consumption ~ ., data = power)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2242603  -114866   -68553    31888  2552368
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)              1.432e+05  6.522e+04   2.195   0.0299 *
## power_index             -1.596e+04  1.287e+04  -1.240   0.2171
## Active.Personnel         1.967e+00  3.659e-01   5.375 3.37e-07 ***
## Armored.Vehicles         3.314e+01  1.994e+01   1.662   0.0989 .
## Helicopters             -1.511e+03  1.165e+03  -1.297   0.1969
## Defense.Budget           3.381e+04  2.848e+03  11.873  < 2e-16 ***
## Navy.Ships              -2.013e+02  6.697e+02  -0.301   0.7642
## Total.Aircraft.Strength -1.321e+02  5.577e+02  -0.237   0.8132
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 514700 on 132 degrees of freedom
## Multiple R-squared:  0.9427, Adjusted R-squared:  0.9396
## F-statistic: 310.1 on 7 and 132 DF,  p-value: < 2.2e-16
```

Predicted OilConsumption = 143178.9 - 1.596343 $\times 10^4$powerindex +1.97ActivePersonnel +33.14Armored-Vehicles - 1510.67Helicopters + 33813.17DefenseBudget + -201.33NavyShips - 132.05TotalAircrafts

R^2 adjusted = 0.9565

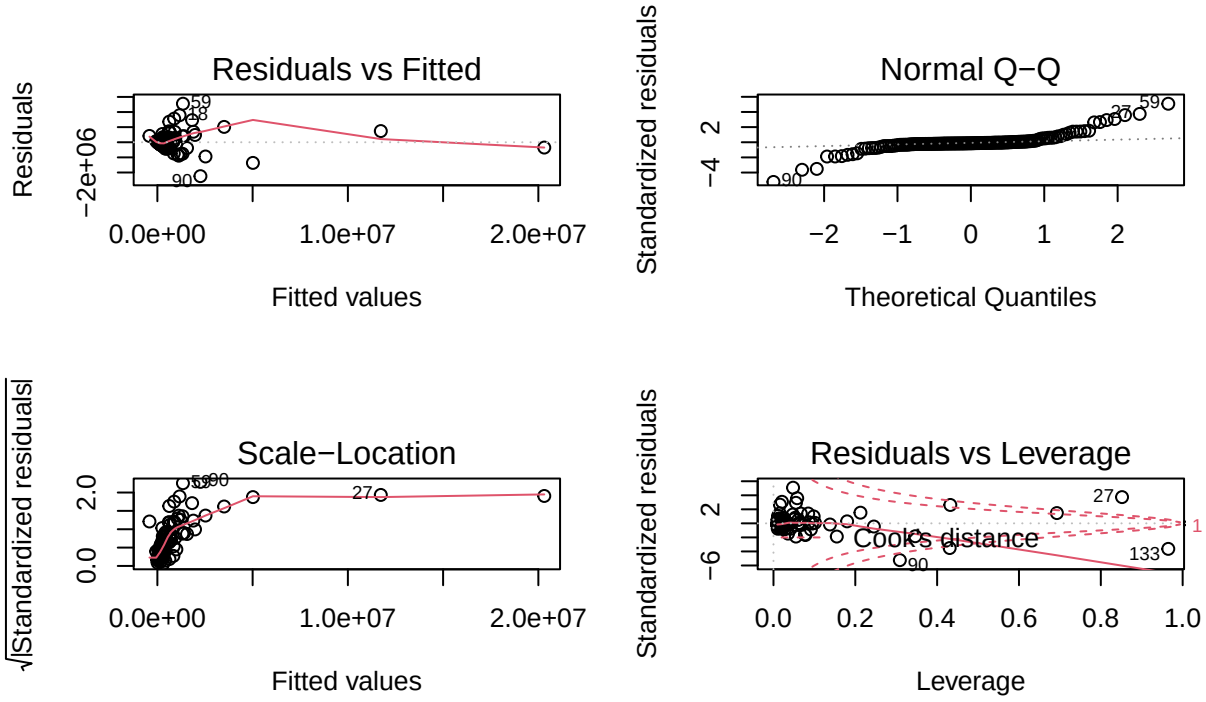
Untransformed Full Model Evaluation Based upon our scatterplot matrix, I know that the correlation coefficient for helicopters, navyships, and aircrafts should be positive. However, our model tells us that the association of these three variables are negative. This result not only goes against the scatterplot matrix, but my intuition as well, as the more aircrafts one has, the more oil is needed in order for them to operate. This suggests that our predictor variables are multicollinear. To check this, I will use the variance inflation factors command in R.

```
vif(m1)
```

```
##             power_index        Active.Personnel        Armored.Vehicles
##                1.097863                5.469301                7.283468
##             Helicopters          Defense.Budget             Navy.Ships
##                172.029152               20.665945                3.365468
## Total.Aircraft.Strength
##                244.103711
```

We find that there is in fact a multicollinearity issue due to the some values being greater than 5.

I shall now check the diagnostic plots to test for linearity, normality of errors, constant variance, and outliers:



Our Residuals vs Fitted Values plot illustrates that the relationship between Oil Consumption and the predictors is non linear. The Normal Q-Q Plot illustrates that our errors are not normal, as the plot appears to be heavily tailed. The Standardized residuals plot display non-constant variance, as there appears to be a slight dip at the very start. Finally, our Residuals vs Leverage plot shows that there is a handful of outliers, as there are points outside [-2,2] standardized residuals and to the right of 2 * (8/140), or 0.1143.

Because there may be an issue with linearity, normality of errors, and constant variance, I shall transform the variables.

Second Model : Power Transformation I shall now perform a power transformation to our first model. Here it is:

```
##         Oil.Consumption             power_index        Active.Personnel
##              0.00000000              -0.09027803              0.21607302
##        Armored.Vehicles             Helicopters          Defense.Budget
##              0.10356583               0.24364423              0.05841230
##             Navy.Ships Total.Aircraft.Strength
##              0.19471267               0.24145419
```

I shall round all values to 0, hence I will apply the log transformation to every variable. The transformed model will look like the following:

```
##
```

```
## Call:
## lm(formula = oil ~ index + active + armored + heli + def + navy +
##     aircrafts)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3.04580 -0.38906  0.02089  0.45533  1.61943
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.44881    0.69070  18.023  < 2e-16 ***
## index       -0.54691    0.15537  -3.520 0.000593 ***
## active      -0.12178    0.04554  -2.674 0.008436 **
## armored     -0.04043    0.07968  -0.507 0.612704
## heli         0.15318    0.15879   0.965 0.336466
## def          0.47130    0.07580   6.218 6.17e-09 ***
## navy         0.14221    0.04376   3.250 0.001465 **
## aircrafts   -0.05535    0.16326  -0.339 0.735143
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6965 on 132 degrees of freedom
## Multiple R-squared:  0.8605, Adjusted R-squared:  0.8531
## F-statistic: 116.4 on 7 and 132 DF,  p-value: < 2.2e-16
```

log(Predicted Oil Consumption) = 12.45 - 0.55log(powerindex) -0.12log(Active Personnel) +-0.04log(ArmoredVehicles) + 0.15log(Helicopters) + 0.47log(DefenseBudget) + 0.14log(NavyShips) - 0.06log(Aircrafts)

Adjusted $R^2 = 0.8578$

Transformed Model Evaluation The correlation coefficient for helicopters is now positive, indicating that the transformation may have been helpful. However, active personnel is now negative, indicating that multicollinearity may still be present. Let us check through calculating the VIFs.

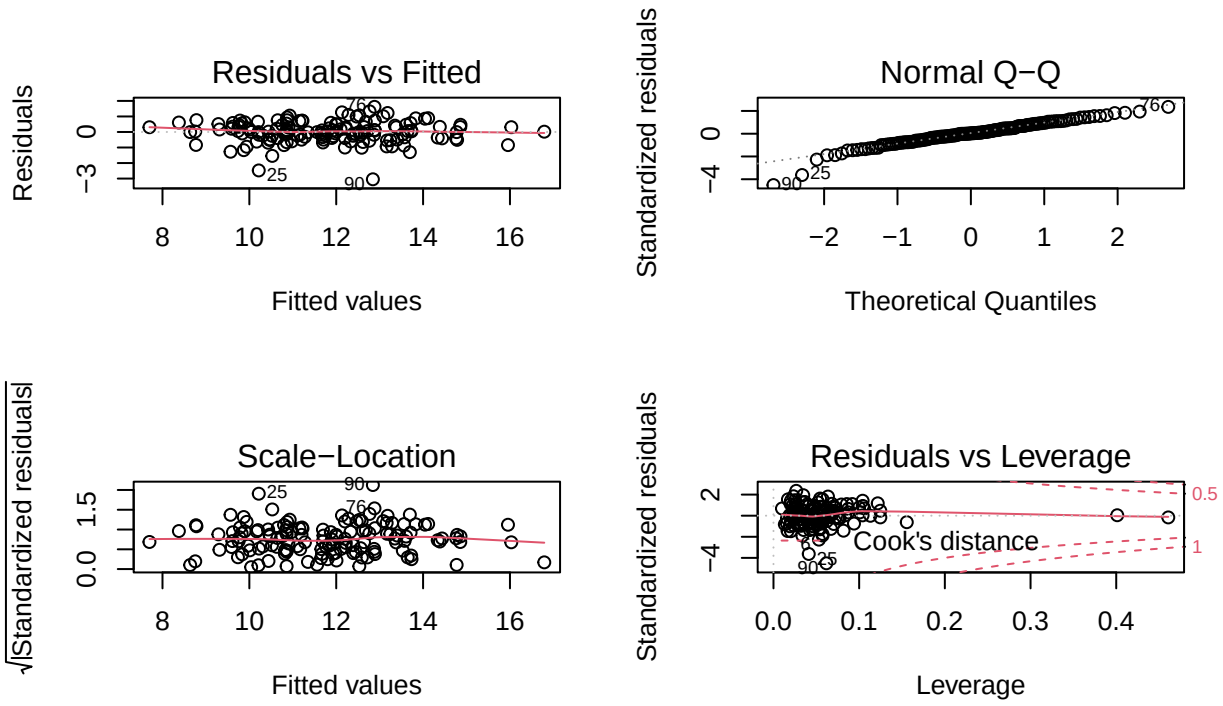
```
vif(power_model)
```

```
##     index    active   armored      heli       def      navy aircrafts
##  9.645689  2.221442  4.669212 17.500465  6.929992  1.963395 23.651468
```

Although the values are significantly smaller, there still appears to be multicollinearity as some values are still much greater than 5.

Let us now analyze the transformed model's diagnostic plots:

```
par(mfrow = c(2, 2))
plot(power_model)
```



All 4 diagnostic plots have significantly improved. The Residuals vs Fitted plot shows that the relationship between the variables is now fairly linear. Although there is still a slight negative slope, it is a significant improvement from the first model's Residuals plot. The Normal Q-Q Plot illustrates that most of the errors are now normal, with just two outliers, observations 25 and 90, not following the linear trend. The Standardized Residuals plot also shows improvement from the first model in that there is constant variance. Finally, the Residuals vs Leverage plots illustrates that there are less outliers compared to the first model. Only observations 25 and 90 are outside of the standardized residuals range of [-2,2]. Additionally, there are less leverage points than the first models, as only two points are to the right of 0.11.
Clearly, the log transformed model is better than the first. However, there is still an issue involving multicollinearity, indicating that there is a need for a reduced model.

Third Model: Reduced Transformed Model To find which variables I should use, I shall do forward stepwise AIC, backward stepwise AIC, forward stepwise AIC, backward stepwise AIC, and use the R^2 adjusted value to choose the best.

```
mint <- lm(oil ~1)
forwardAIC <- step(mint, scope = list(lower = ~1, upper = power_model), direction = "forward")

backwardAIC <- step(power_model, direction = "backward")

forwardBIC <- step(mint, scope = list(lower = ~1, upper = power_model), direction = "forward", k=log(140

backwardBIC <- step(power_model, direction = "backward", k = log(140))
```

The Four tests did not agree on the same model:
- Forwards AIC and Backwards AIC chose: DefenseBudget, Index, NavyShips, ActivePersonnel, Tanks, Heli
- Forwards BIC and Backwards BIC chose: DefenseBudget, Index, NavyShips, ActivePersonnel

To choose between the two we shall look at the adjusted R^2 values and conduct a partial F-test

```
x1 <- lm(oil ~ def + index + navy + active + aircrafts + heli)
summary(x1)$adj.r.squared
```

```
## [1] 0.8539641
```

```
x2 <- lm(oil ~ def + index + navy + active)
summary(x2)$adj.r.squared
```

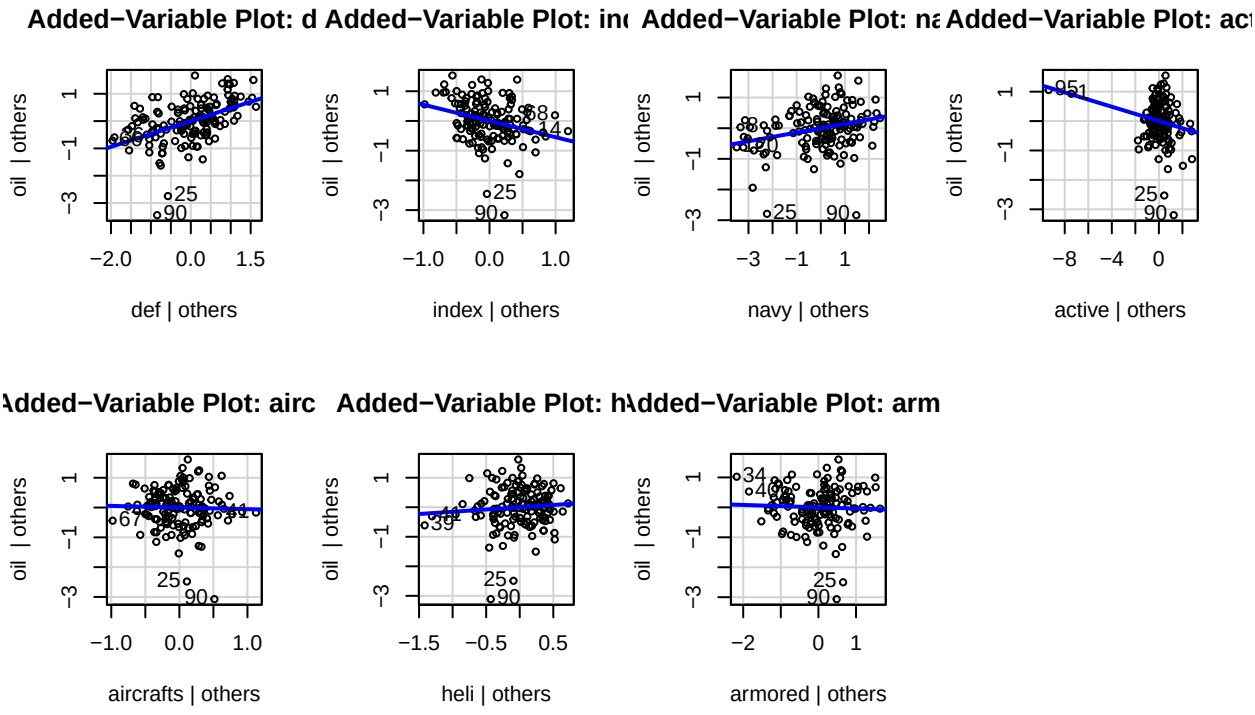
```
## [1] 0.8545772
```

The adjusted R-squared values says the model with just defense budget, power index, navy ships, and active personnel is best. Let us now conduct a partial F-test

```
## Analysis of Variance Table
##
## Model 1: oil ~ def + index + navy + active + aircrafts + heli
## Model 2: oil ~ def + index + navy + active
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    133 64.161
## 2    135 64.852 -2  -0.69139 0.7166 0.4903
```

The p-value is greater than 0.05, so we fail to reject the null hypothesis and accept the reduced model. This is in line with what the adjusted R^2 values told us.

Let us now look at the added variable plots to confirm that this reduced model is best:

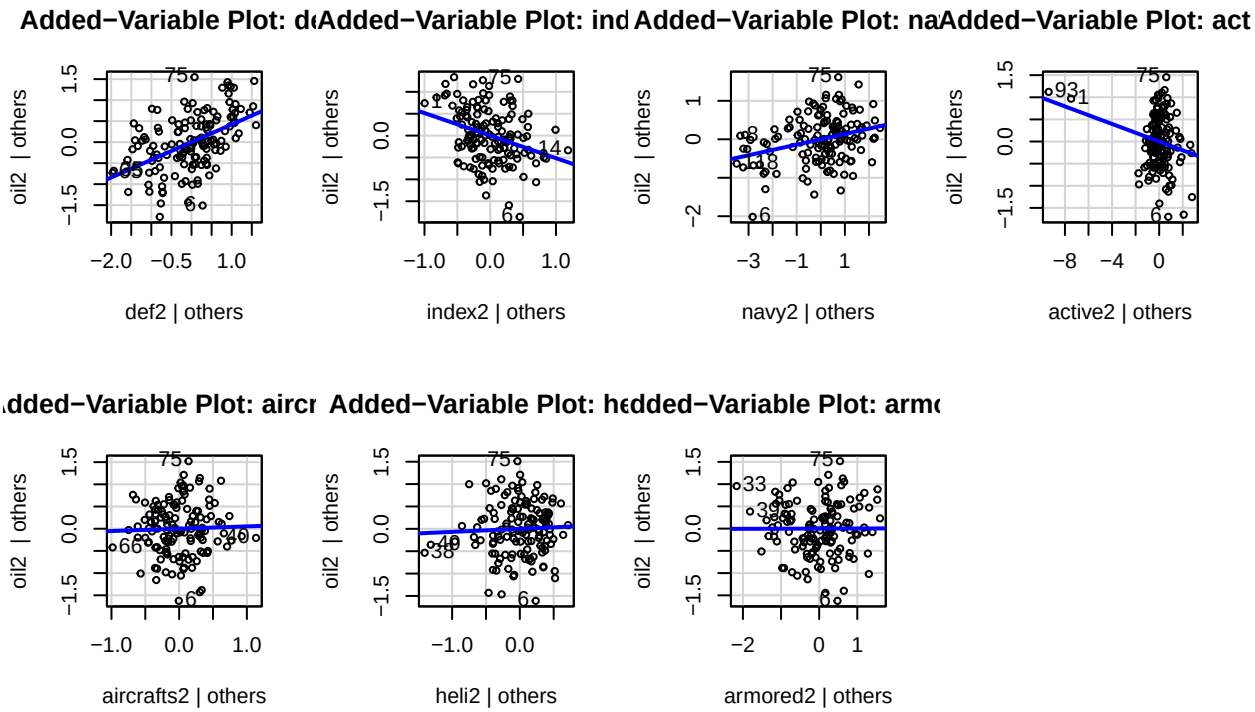


The added variable plots align with the conclusion that the reduced model is best. This is because the slopes of aircrafts, helicopters, and armored vehicles are non-significant. The plots also illustrate that observations 25 and 90 are problematic, just as our diagonistic plots had shown as well. They heavily skew the graphs as they are significantly away from our cluster of points. For this reason, I shall see how these added variable plots look like with these points removed.

Removing outliers:

```
power3 <- power2[-c(25,90),]
oil2 <- log(power3$Oil.Consumption)
index2 <- log(power3$power_index)
active2 <- log(power3$Active.Personnel + 1) # add plus ones because log(0) is undefined
armored2 <- log(power3$Armored.Vehicles)
heli2 <- log(power3$Helicopters + 1)
def2 <- log(power3$Defense.Budget)
navy2 <- log(power3$Navy.Ships + 1)
aircrafts2 <- log(power3$Total.Aircraft.Strength + 1)
outlier_model <- lm(oil2 ~ index2 + active2 + armored2 + heli2 + def2 + navy2 + aircrafts2)
```

The new added variable plots:



Removing the outliers did not seem to significantly affect the added variable plots, so I will still go with the reduced model without removing the two outliers. Here is the equation of our reduced model:

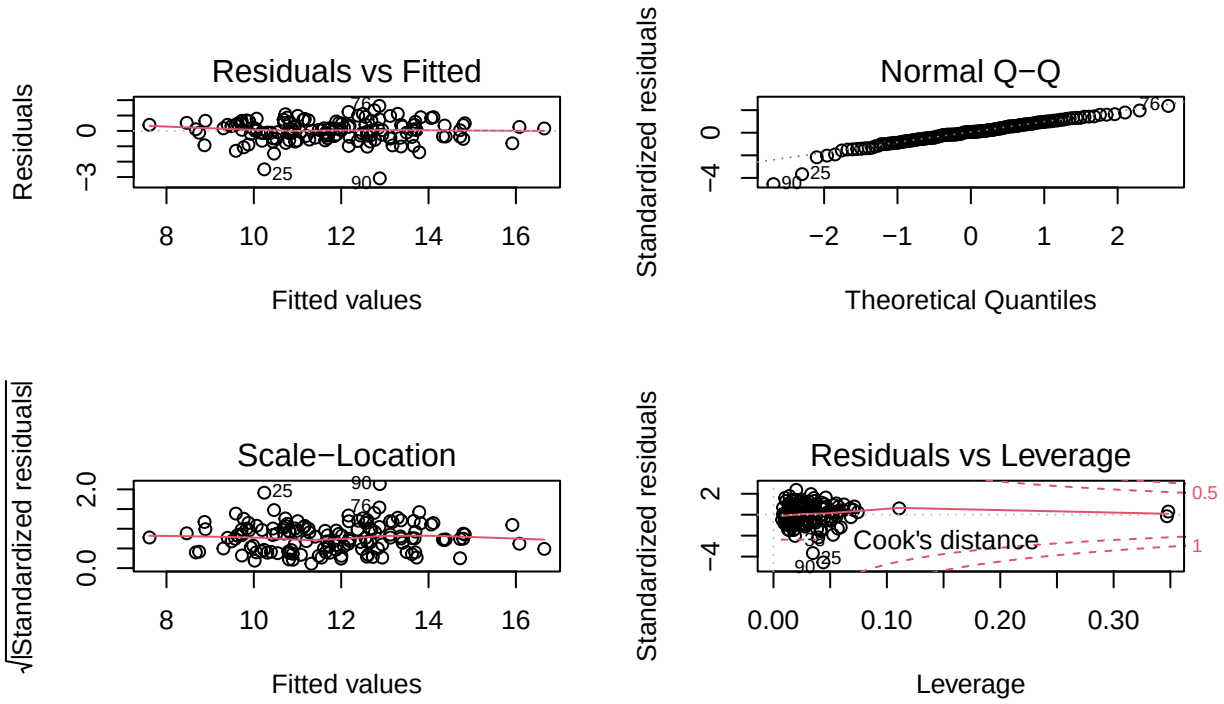
```
##
## Call:
## lm(formula = oil ~ def + index + navy + active)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3.08518 -0.37638  0.04957  0.46621  1.62628
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.33928    0.45534  27.099  < 2e-16 ***
## def          0.47235    0.07046   6.704 5.07e-10 ***
## index       -0.57811    0.12977  -4.455 1.74e-05 ***
## navy         0.15071    0.04257   3.540 0.000549 ***
## active      -0.10957    0.04243  -2.582 0.010878 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6931 on 135 degrees of freedom
## Multiple R-squared:  0.8588, Adjusted R-squared:  0.8546
## F-statistic: 205.2 on 4 and 135 DF,  p-value: < 2.2e-16
```

log(Predicted Oil Consumption) = 12.34 + 0.47log(DefenseBudget) -0.58log(power_index) +0.15log(NavyShips) - 0.11log(ActivePersonnel)

Adjusted $R^2 = 0.8546$

Reduced Transformed Model Evaluation Let us check the diagonistic plots:



All our assumptions are satisfied. The Residuals vs Fitted plot illustrates a linear relationship, the normal Q-Q plot shows a normality of errors, and the standardized residual plot shows standardized residuals. The only significant outliers are points 25 and 90 according to the leverage plot.

Now let us check for multicollinearity:

```
vif(x2)
```

```
##      def    index     navy   active
## 6.047322 6.794439 1.876587 1.947589
```

The VIFs are still a little high, but they are significantly better than the VIFs of model one and two.

Final Model Because all the assumptions of linearity, normality of errors, and constance variance are satisfied, and multicollinearity is significantly reduced, our reduced transformed model is our final model.

log(Predicted Oil Consumption) = 12.34 + 0.47log(DefenseBudget) -0.58log(power_index) +0.15log(NavyShips) - 0.11log(ActivePersonnel)

- If Power_index, Navyships, and Active Personnel are held constant, than a 1% increase in Defense Budget results in a 0.47% increase in Oil Consumption
- If Defense Budget, Navyships, and Active Personnel are held constant, then a 1% increase in Power_index will result in a 0.58% decrease in Oil Consumption
- If Defense Budget, Power_index, and Active Personnel are held constant, then a 1% increase in Navy Ships will result in a 0.15% increase in Oil Consumption
- If Defense Budget, Power_index, and Navy Ships are held constant, then a 1% increase in Active Personnel will result in a 0.11% decrease in Oil Consumption.

Discussion According to an article published in ThePrint, the US military is the largest consumer of fossil fuels in the world (Weir et al., 2021). Another study conducted in 2019 found that the US military's use would "make it the 47th largest emitter of greenhouse gases in the world" (Belcher et.al, 2019). However, we cannot confirm exactly how much each nation's military consumes, as in 1997, the US won a "military exemption" from reporting under the Kyoto climate accord. This military exemption is granted under the grounds of "national security", according to the Military Emissions Gap. For this reason, I think that our final reduced transformed model can give many insights into how a nation's military impacts a nation's oil use.

I believe that the final reduced transformed model makes a lot of sense in the real world. The reason why I believe this is because most of the correlation coefficients align with intuition. For instance, one would expect that if a military were to increase their budget, then more money would go into building, transporting, and inventing projects which require a great deal of energy to construct. Hence, oil consumption would increase. The same reasoning can be said for the power_index and navy ships variables as well.

Something which does not make sense, however, is the active personnel correlation coefficient being negative. I would expect in the real world that the more active military members, the more oil would be needed in order to transport, house, and accommodate for them. Additionally, I am suprised that the aircrafts variable was not found to be significant. I would expect aircrafts to be the most important factor of them all, as commercial airplanes and jets contribute around 3 percent of the nations total greenhouse gas production (Overton, 2019).

Limitations and Future Improvements Our model had to overcome many limitations. The main one was the one stated above: that militaries do not have to report their use. For this reason, the oil consumption from our data set was the total oil consumption of a nation, not the oil consumption from the military alone. For this reason, there could be many outliers, as countries may use tons of oil while having an inadequate military. However, for the most part, there is an association between oil use and military strength.

Areas where the final model could improve is by analyzing a categorical variable. For instance, in the original data frame, there is a column called Aircraft Carriers. Aircraft carriers are a true flagship of the most powerful militaries in the world. In fact, only seven countries have just one, and the US is the only nation with more than 2. So the jets you see taking off from Top Gun can only be done from seven countries. If I wanted to have a categorical variable, I could make having an aircraft carrier a categorical variable and seeing if this significantly affects the oil consumption of a nation.

Sources

Belcher, O, Bigger, P, Neimark, B, Kennelly, C. Hidden carbon costs of the "everywhere war": Logistics, geopolitical ecology, and the carbon boot-print of the US military. Trans Inst Br Geogr. 2020; 45: 65– 80. https://doi.org/10.1111/tran.12319

Environmental and Energy Study Institute (EESI). "Issue Brief: The Growth in Greenhouse Gas Emissions from Commercial Aviation (2019, Revised 2022)." EESI, https://www.eesi.org/papers/view/fact-sheet-the-growth-in-greenhouse-gas-emissions-from-commercial-aviation

Kanawattanachai, Prasert. "Military Power by Country 2022." Kaggle, 20 Feb. 2022, https://www.kaggle.com/datasets/prasertk/military-power-by-country-2022.

Weir, Doug. "US Defense Largest Consumer of Fossil Fuels - How World's Militaries Hide Carbon Emissions." ThePrint, 14 Nov. 2021, https://theprint.in/world/us-defense-largest-consumer-of-fossil-fuels-how-worlds-militaries-hide-carbon-emissions/765688/.

"2023 Military Strength Ranking." Global Firepower - World Military Strength, https://www.globalfirepower.com/countries-listing.php.