STAT6020-Assignment4-TV

Exploring the relationships that contribute to CO2 emissions and the impact that has on renewable energy contributions

Abstract:

Our study has been commissioned to explore the world bank group climate change data to support world leaders tasked with considering and implementing policy to safeguard the planet's future at the upcoming 2021 UN Climate Change conference.

The purpose of this report is to deliver a statistical analysis, interpretation and discussion of findings with relation to the significant predictors of CO2 emissions and relationship CO2 has with renewable energy consumption as a percentage of total country energy consumption.

We will perform a multiple linear regression analysis to determine the relationships that can predict high CO2 emissions for 163 World Bank member countries. Additionally, we will engage in regularisation and variable selection to determine the most important variables for consideration within our model. Lastly, we will explore a principal component analysis to understand the direction and magnitude of our significant variables.

In our report we've found significant relationships between CO2 emissions and explanatory variables supporting rejection of a null hypothesis that selected climate change explanatory variables have no relationship with CO2 emissions. Furthermore, we've identified the strongest influences on CO2 emissions as a subset of selected variables of interest and were unable to reject the null hypothesis that foreign direct investment has no relationship with CO2 emissions.

Our results pave way for world leaders to consider exactly what country-level features should be considered when developing global policy and to what magnitude selected climate change explanatory variables impact CO2 emissions to weight policy and levies accordingly.

1 Introduction:

Climate change is a serious global threat and demands an urgent global response {Norling, 2008 #196}. As we see the population grow and the demand for production throughout the world rise, the topic of climate change intensifies as the 21st Century's greatest environmental and political challenge. CO2 emissions in 2008 were forecasted to increase by 45% by 2030 {International Energy Agency., 2008 #197} and with the hopes of a united geopolitical landscape in recent years having been decimated by Covid-19, increasing CO2 emissions remain as great a threat to humanity as ever before.

This report will explore, analyse and interpret world bank climate change data to help map a course of action for governmental stakeholders to incorporate as part of a united Earth plan to tackle climate change and reduce CO2 emissions.

Studies have shown that human emissions of carbon dioxide and other greenhouse gases are the primary drivers of climate change {Change, 2007 #200}, our primary objective in this report will be to analyse and understand the specific drivers of high CO2 emissions, including what relationship renewable energy consumption has with CO2 emission production.

A secondary objective of this report will be to determine the impact of foreign direct investment (FDI) on countries across the world. With so many reasons at play for foreign direct investment, there are mixed results in studies showing that in some scenarios FDI inflows can reduce CO2 emissions {Zhang, 2016 #201} whereas in other scenarios FDI increases carbon emissions intensity {Hu, 2021 #202} contributing to greater greenhouse gas (GHG) emissions.

2 Data:

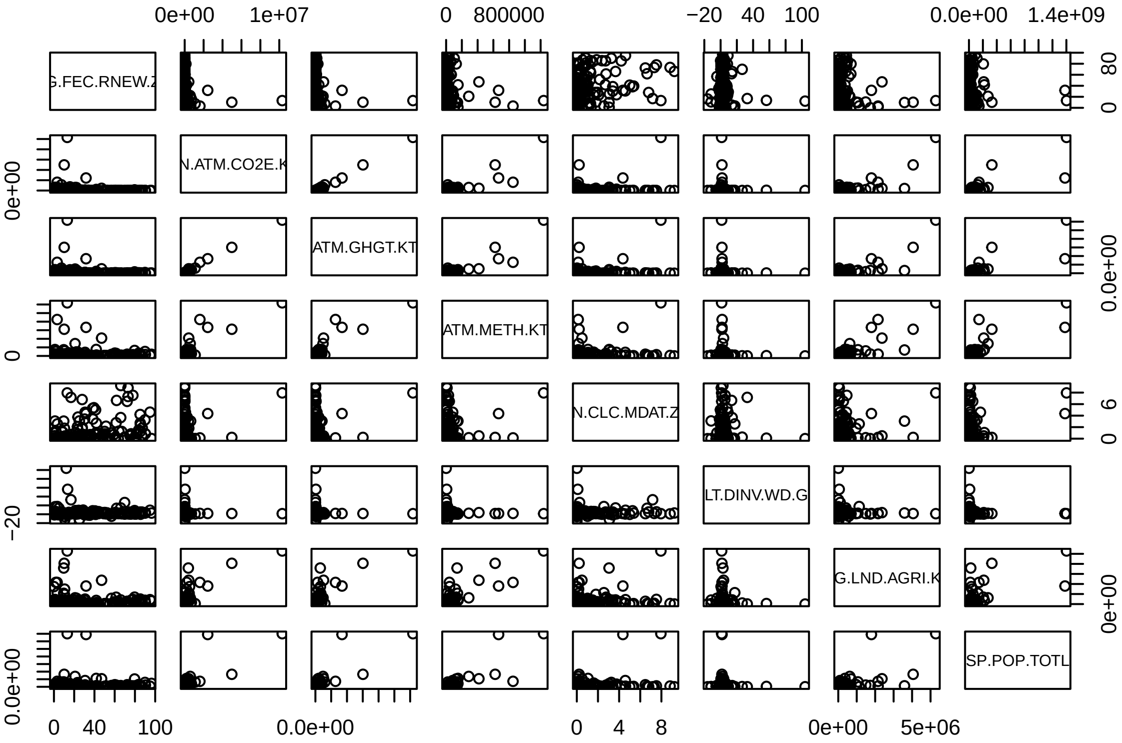
```
wbdata <- read.csv("wbcc_bc.csv") # reading in data

wb = wbdata %>%
  select(country, EG.FEC.RNEW.ZS, EN.ATM.CO2E.KT, EN.ATM.GHGT.KT.CE, EN.ATM.METH.KT.CE, EN.CLC.MDAT.ZS,
wb = na.omit(wb)
wbnum = subset(wb, select = -country) # removing non non-numeric
```

```
summary(wbnum)
```

```
##  EG.FEC.RNEW.ZS    EN.ATM.CO2E.KT     EN.ATM.GHGT.KT.CE  EN.ATM.METH.KT.CE
##  Min.   : 0.0157   Min.   :      80   Min.   :     110   Min.   :     20
##  1st Qu.:11.0890   1st Qu.:    4190   1st Qu.:   12765   1st Qu.:   3635
##  Median :25.7461   Median :   16580   Median :   45050   Median :  11320
##  Mean   :33.8610   Mean   :  202771   Mean   :  276547   Mean   :  48976
##  3rd Qu.:52.4332   3rd Qu.:   81125   3rd Qu.:  121245   3rd Qu.:  40805
##  Max.   :96.3837   Max.   :10313460   Max.   :12355240   Max.   :1238630
##  EN.CLC.MDAT.ZS    BX.KLT.DINV.WD.GD.ZS AG.LND.AGRI.K2     SP.POP.TOTL
##  Min.   :0.00000   Min.   :-16.062      Min.   :     16   Min.   :5.919e+04
##  1st Qu.:0.02208   1st Qu.:  1.264      1st Qu.:  16646   1st Qu.:3.377e+06
##  Median :0.26047   Median :  2.496      Median :  52960   Median :1.070e+07
##  Mean   :1.18588   Mean   :  4.057      Mean   : 290407   Mean   :4.671e+07
##  3rd Qu.:1.27563   3rd Qu.:  4.213      3rd Qu.: 260225   3rd Qu.:3.452e+07
##  Max.   :9.22659   Max.   :103.933      Max.   :5285287   Max.   :1.402e+09
```

```
pairs(wbnum)
```



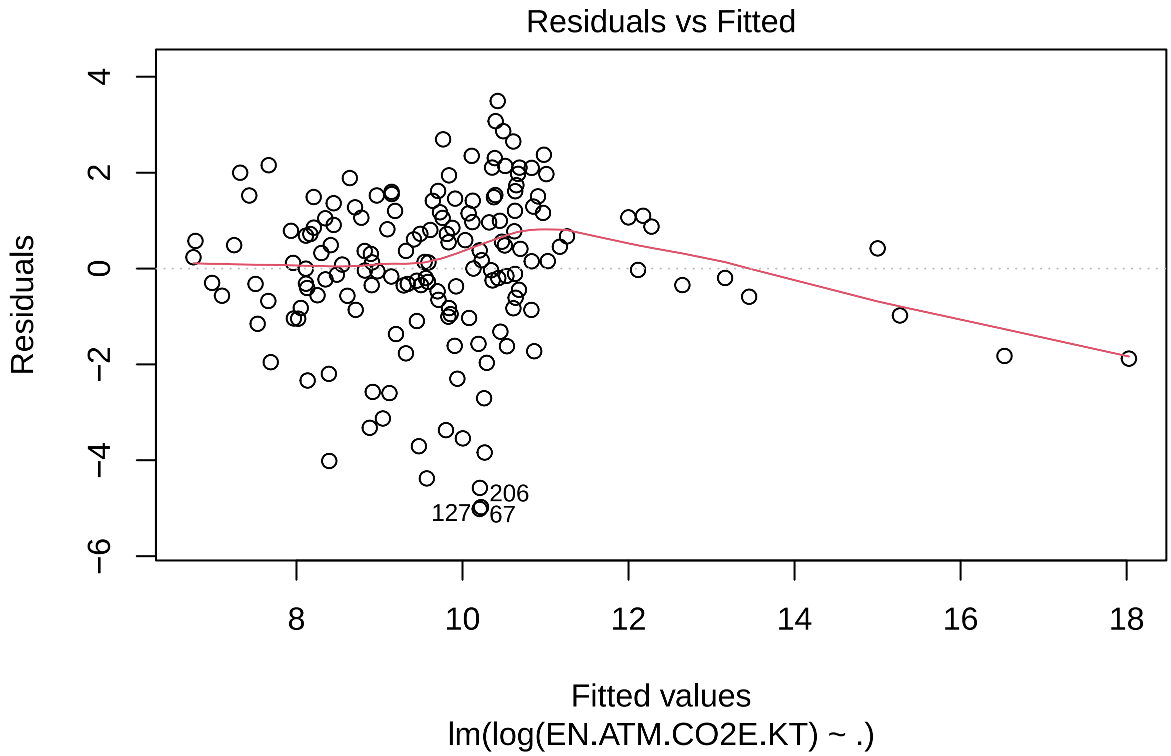
A brief analysis of pairs() and summary () shows that there are great variations in data distribution and means, including a few outliers present across multiple variables. Due to the scale of our response variable CO2 emissions, in contrast to the explanatory variables, we will be looking to perform a log transformation to improve the distribution.

3 Methods:

3.1 Multiple Linear Regression

We will first begin our statistical analyses by determining the significance of relationships between CO2 emissions and our selected predictors using a multiple linear regression model, then performing regularisation and variable selection to determine the best set of variables to use within our model.

```r
# multiple linear regression model
wbnum.lm = lm(log(EN.ATM.CO2E.KT)~., data=wbnum)
plot(wbnum.lm, which=1)
```



Here we have plotted the residuals for the linear model with log(EN.ATM.CO2E.KT) as our response and observe the smoothing line skew towards the aforementioned outliers. Our explanatory variables exhibit a weak but somewhat linear relationship.

```
summary(wbnum.lm)
```

```
##
## Call:
## lm(formula = log(EN.ATM.CO2E.KT) ~ ., data = wbnum)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.0137 -0.7491  0.1371  1.0834  3.4925
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)           1.064e+01  2.337e-01  45.539  < 2e-16 ***
## EG.FEC.RNEW.ZS       -3.193e-02  5.222e-03  -6.114 7.58e-09 ***
## EN.ATM.GHGT.KT.CE    -4.545e-07  3.009e-07  -1.510  0.13298
## EN.ATM.METH.KT.CE     4.005e-06  2.560e-06   1.564  0.11982
## EN.CLC.MDAT.ZS       -1.884e-01  7.338e-02  -2.567  0.01120 *
## BX.KLT.DINV.WD.GD.ZS -2.013e-02  1.294e-02  -1.556  0.12179
## AG.LND.AGRI.K2        9.417e-07  3.548e-07   2.654  0.00878 **
## SP.POP.TOTL           3.569e-09  1.660e-09   2.150  0.03309 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 1.669 on 155 degrees of freedom
## Multiple R-squared:  0.4894, Adjusted R-squared:  0.4664
## F-statistic: 21.23 on 7 and 155 DF,  p-value: < 2.2e-16
```

```
mean(wbnum.lm$residuals^2) # MSE
```

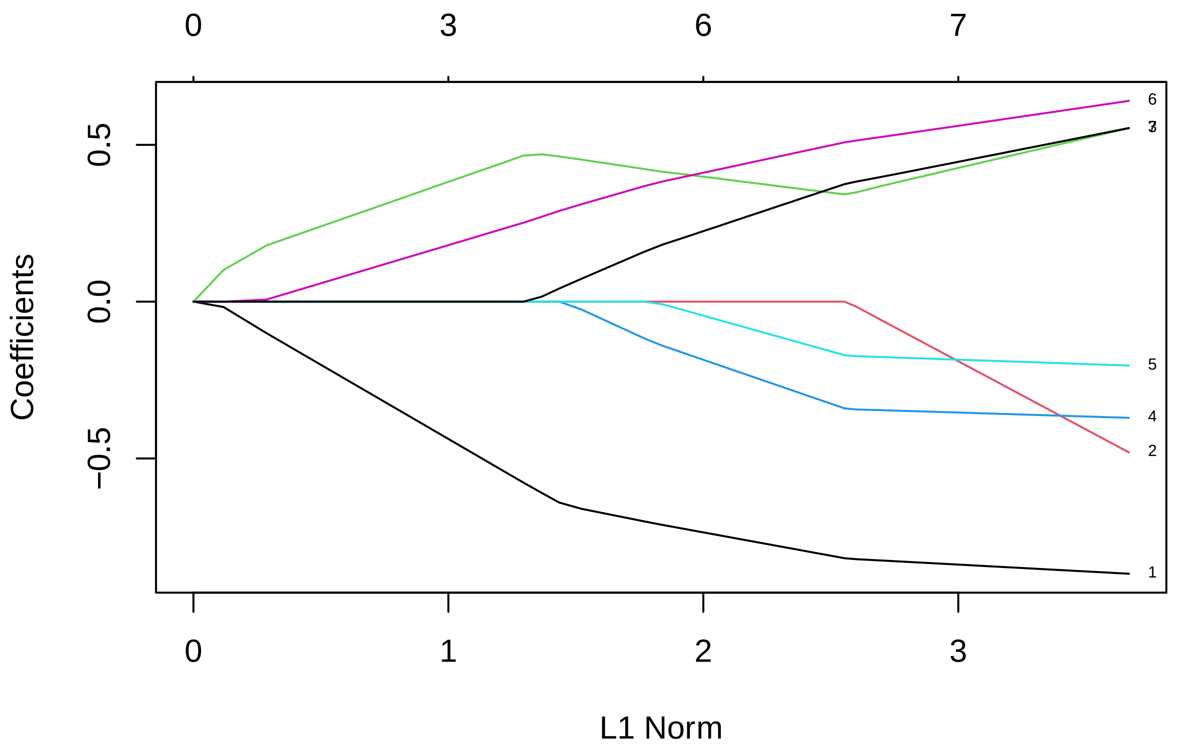
```
## [1] 2.649198
```

In summary of our linear model, we report a multiple R-squared statistic of 0.4894 where the model accounts for approximately half of the variance in the response variable. We appoint a significance measure of p value $= \leq .05$ and therefore identify 4 significant predictors of the response variable CO2 emissions as; renewable energy consumption, natural disasters and extreme temperatures, agricultural land size and total population. Our RSE is 1.669 and we calculate a MSE of 2.649.

3.2 LASSO regression

Now, to improve upon performance of our model we will undertake LASSO regression and introduce regularisation using a technique known as shrinkage. We have scaled the X variables within the function to clearly visualise which variables converge to 0 last.

```
X = model.matrix(wbnum.lm)[,-1]
Y = log(wbnum$EN.ATM.CO2E.KT)
lasso.wb = glmnet(scale(X), Y, alpha=1) #scaled X
plot(lasso.wb,label=TRUE)
```



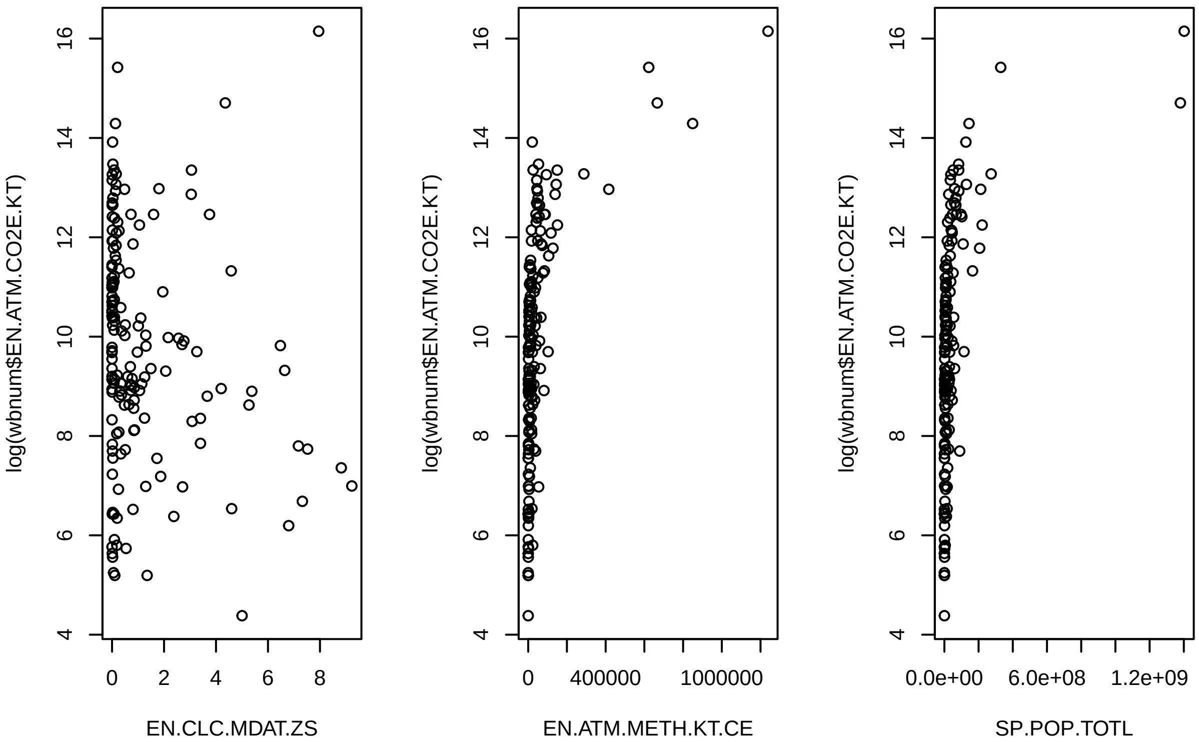
Reading this plot from right to left, a higher penalty term λ value is being applied eventually reducing all coefficients to 0, we can see the number of variables that remain with the increasing value of λ applied on the top axis.

This variable convergence to 0 depicts the importance of the variables in our model, given that variables 3, 1 an 6 (methane emissions, renewable energy consumption and agricultural land mass) all converge to 0 last, we can conclude these predictors are of the most importance for our model.

The methane emissions explanatory variable is particularly interesting being the last variable to converge to 0 and increasing in coefficient value temporarily as the value of λ increased. Methane emissions were not identified as being significant in our linear model, whereas natural disasters and extreme temperatures and population were. In contrast, methane is one of the most important variables selected as part of our LASSO regression.

What could be causing this result? We explore further by observing the plots side by side.

```
par(mfrow=c(1,3))
plot(log(wbnum$EN.ATM.CO2E.KT)~EN.CLC.MDAT.ZS , data=wbnum)
plot(log(wbnum$EN.ATM.CO2E.KT)~EN.ATM.METH.KT.CE , data=wbnum)
plot(log(wbnum$EN.ATM.CO2E.KT)~SP.POP.TOTL , data=wbnum)
```

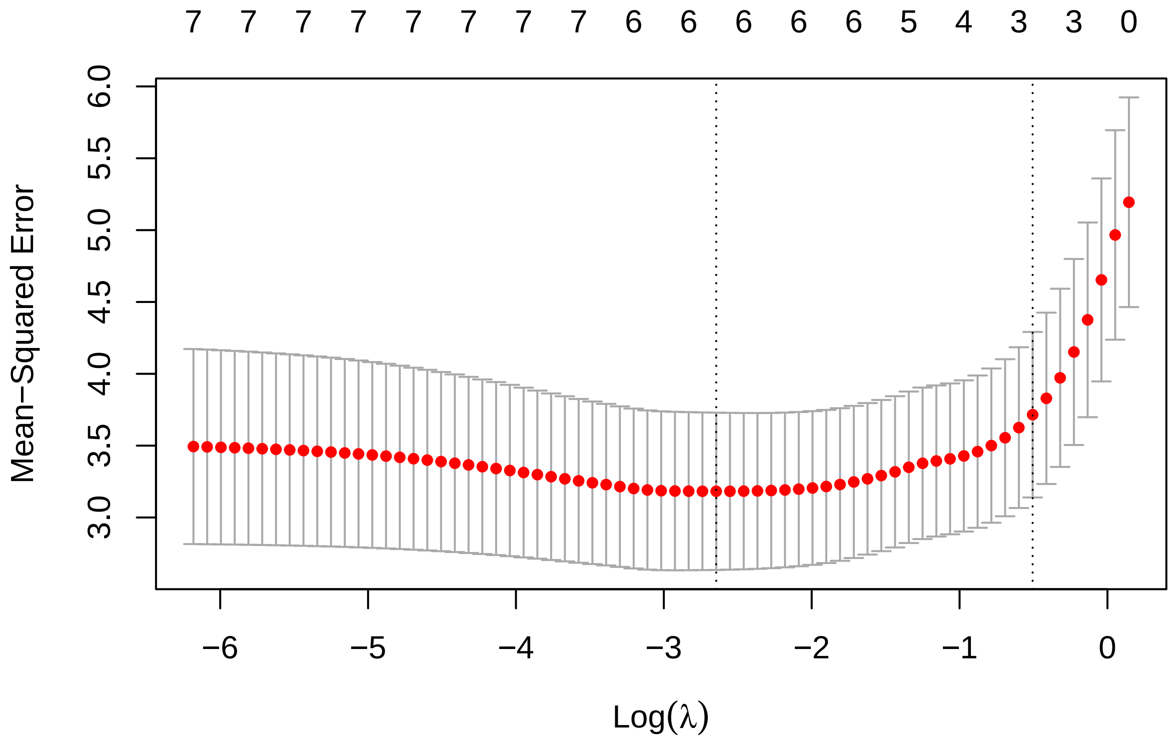


In exploration of this difference, we have plotted natural disasters and extreme temperatures and methane emissions individually against log(EN.ATM.CO2E.KT) and observe a potential risk for multicollinearity within our linear model. Given the penalty term λ applies to each variable individually on a scale increasing in value until all coefficients converge to zero, we are confident reporting methane as one of the important variables proposed by the model having observed it's convergence to zero last.

As another method to confidently determine which variables are the best fit for our model, we will implement a default 10-fold cross-validation on our lasso regularisation model and use a measure of minimum mean cross-validated error to determine best fit.

3.3 Regularisation & variable selection with cross-validated LASSO

```
#cross-validated lasso
cv_lasso = cv.glmnet(X,Y,alpha=1)
plot(cv_lasso)
```



In this plot from left to right we observe the value of $log(\lambda)$ increasing and therefore reducing the number of coefficients in the cross-validated lasso regression model eventually to zero. Each red dot represents an average mean square error calculated for each increasing value of lambda. We are looking to find the minimum mean cross-validated error and optimal value of lambda.

The optimal value of lambda where the mean squared error is the smallest is denoted by the dotted line on the right and corresponds to a $log(\lambda)$ of -2.645662.

```
min(cv_lasso$cvm) # minimum mean cross-validated error
```

```
## [1] 3.182281
```

```
log(cv_lasso$lambda.min) # value of lambda that gives minimum cvm
```

```
## [1] -2.645662
```

We observe no significant change in the MSE between the models of 3.14 for our cross-validation LASSO model and 2.65 for the linear model.

With this penalty term applied, our coefficient variable selection and values is as follows:

```
coef(cv_lasso)
```

```
## 8 x 1 sparse Matrix of class "dgCMatrix"
##                                  s1
## (Intercept)           1.008297e+01
## EG.FEC.RNEW.ZS       -1.514773e-02
## EN.ATM.GHGT.KT.CE     .
## EN.ATM.METH.KT.CE     2.604800e-06
## EN.CLC.MDAT.ZS        .
## BX.KLT.DINV.WD.GD.ZS  .
## AG.LND.AGRI.K2        2.429388e-07
## SP.POP.TOTL           .
```

We note here that our cross-validated coefficients compared with our linear regression model are smaller in magnitude as a result of the optimisation. We also observe variable selection taking place, with renewable energy consumption, methane emissions and agricultural land mass selected as the optimal variables by our cross-validation LASSO model. This result is different from the linear model which included natural disasters and extreme temperatures, and population as significant in relation to the response variable.

3.4 Principal component analysis

To further visually examine these relationships, we will explore the data using principal component analysis, observing the direction and magnitude of the selected variables in a plot of our two principal components that account for the majority of variance in our model.

```
wbnum2 = subset(wb, select=c(EN.ATM.CO2E.KT, EG.FEC.RNEW.ZS, EN.ATM.METH.KT.CE, AG.LND.AGRI.K2))
wbpca = prcomp(wbnum2, scale=TRUE) # PCA analysis scaled variables
summary(wbpca)
```

```
## Importance of components:
##                           PC1    PC2     PC3     PC4
## Standard deviation     1.6369 0.9828 0.47866 0.35402
## Proportion of Variance 0.6699 0.2415 0.05728 0.03133
## Cumulative Proportion  0.6699 0.9114 0.96867 1.00000
```

```
wbpca$sdev^2/sum(wbpca$sdev^2) # % of variance in each loading vector
```

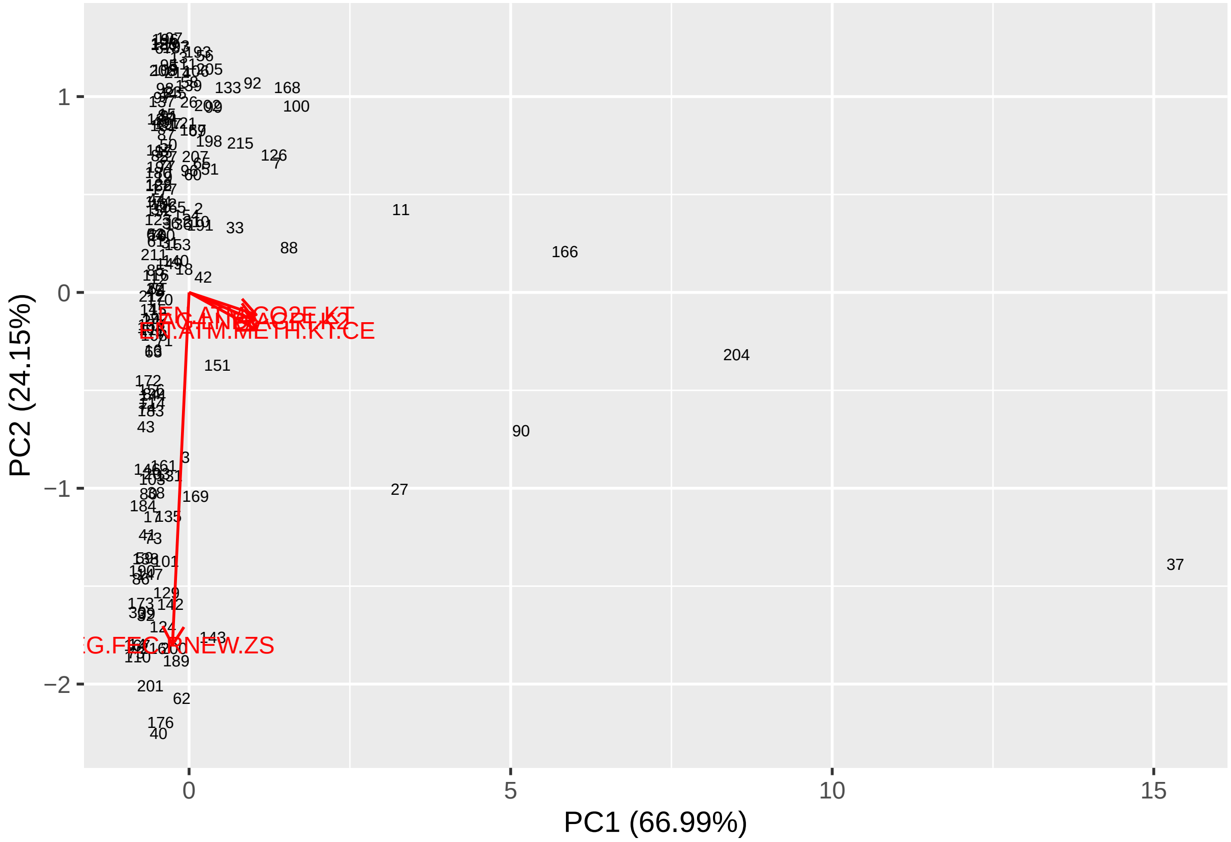
```
## [1] 0.66989084 0.24149671 0.05727975 0.03133270
```

```
wbpca$rotation # rotated observations coefficients
```

```
##                          PC1         PC2          PC3         PC4
## EN.ATM.CO2E.KT     0.5730995 -0.06260301  0.559196480  0.59576599
## EG.FEC.RNEW.ZS    -0.1423573 -0.98938527  0.004407259  0.02884001
## EN.ATM.METH.KT.CE  0.5803581 -0.10516143  0.220960169 -0.77672528
## AG.LND.AGRI.K2     0.5607814 -0.07834997 -0.799034716  0.20230911
```

Here we observe the standard deviation for each of the loading vectors and have calculated the percentages for each of the PC vectors. Additionally, we have output the coefficients for the rotated PC vectors.

```
autoplot(wbpca,data=wbnum, loadings = TRUE, loadings.label = TRUE, scale = 0,
         shape = FALSE, label.size = 2, loadings.label.size=3,
         colour = "black")
```



We have plotted PC 1 and PC 2 which together make up 91.14% of variance in the data. It is interesting to note the similar direction and magnitude of all variables except renewable energy consumption.

By investigating specific outliers denoted in our plot by the respective country's ID, we can understand how invested the largest producers of emissions are placed in terms of their renewable energy consumption.

```
summary(wbnum2$EG.FEC.RNEW.ZS)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0157 11.0890 25.7461 33.8610 52.4332 96.3837
```

By printing a summary of renewable energy, we can see the mean value is 25.7461.

```
# outlier observations from PC plot with high CO2 values
wbdata[37,c("country","EG.FEC.RNEW.ZS", "EN.ATM.CO2E.KT")]
```

```
##    country EG.FEC.RNEW.ZS EN.ATM.CO2E.KT
## 37   China        13.1238       10313460
```

```
wbdata[204,c("country","EG.FEC.RNEW.ZS", "EN.ATM.CO2E.KT")]
```

```
##           country EG.FEC.RNEW.ZS EN.ATM.CO2E.KT
## 204 United States        10.1072        4981300
```

```
wbdata[166,c("country","EG.FEC.RNEW.ZS", "EN.ATM.CO2E.KT")]
```

```
##                country EG.FEC.RNEW.ZS EN.ATM.CO2E.KT
## 166 Russian Federation         3.1805        1607550
```

```
wbdata[90,c("country","EG.FEC.RNEW.ZS", "EN.ATM.CO2E.KT")]
```

```
##    country EG.FEC.RNEW.ZS EN.ATM.CO2E.KT
## 90   India        31.6892        2434520
```

China, United States, Russia and India reside far beyond the mean of our PCA analysis scaled variables. An interpretation can be applied where for a small set of those observations printed above, a significant percentage of CO2 emissions are being produced by a few countries all of which have renewable energy consumption values well below the mean.

In summary of our statistical analyses, we have produced a multiple linear regression model, performed regularisation and variable selection with cross-validated lasso regression to find a set of explanatory variables that feature significant relationships with the response variable CO2 emissions. Finally, we used principal component analysis to explore the direction and magnitude of our significant variables to shed more insight how much the world's most polluting countries support renewable energy consumption.

4 Further Discussion:

Our findings clearly indicate that the most significant influences on CO2 from the subset of climate change data we've analysed are renewable energy consumption, methane emissions and agricultural land mass.

In 3.1 Multiple Linear Regression, we saw that for every 1 value increase in renewable energy consumption, there was a significant (p value ≤ 0.05) negative relationship with CO2 emissions, a t-statistic of -6.114 showed the standard error was sufficiently far enough away from the coefficient to confidently support a view that a higher renewable energy consumption as a percentage of total energy consumption resulted in lower CO2 emissions.

Interestingly, direct foreign investment was not found to be a significant explanatory variable of CO2 emissions validated by our inability to reject the null hypothesis in all three of our models 3.1, 3.3 and 3.3 with p-values of $\geq .05$.

Our principal component analysis painted a concerning picture that spotlighted the disheartening lack of renewable energy efforts by those few select countries producing the majority of the worlds CO2 emissions. This raises the question of whether the largest economies in the world are truly challenged by climate change or whether they are more interested in profit over sustainable solutions.

5 Conclusions:

In conclusion, the information found in this report can be used for consideration of what country-level features should be chosen when introducing weights for climate-change measures such as production caps, taxes or levies.

World leaders - especially those functioning mostly on renewable energy - should apply more pressure to the handful of countries which are responsible for extreme levels of pollution, demanding they invest more in renewable energy solutions to offset their damaging actions to the planet.

The United States, China, India and Russia have large populations with large agricultural land masses producing lots of emissions which is worsened by the fact of their low renewable energy consumption rates as a percentage of total energy consumption. Statistically, these country profiles are responsible for the modern world's catastrophic climate change situation and should be the focus of UN climate change leaders.

6 References:

Norling, J. (2008). The economics of climate change. Australian Planner, 45(4), 20-23. doi:10.1080/07293682.2008.10753385

International Energy Agency., & OECD iLibrary. (2008). World energy outlook 2008. Retrieved from https://ezproxy.newcastle.edu.au/login?url=https://doi.org/10.1787/weo-2008-en

Change, I. C. (2007). The physical science basis. Contribution of working group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change, 996.

Zhang, C., & Zhou, X. (2016). Does foreign direct investment lead to lower CO2 emissions? Evidence from a regional analysis in China. Renewable and Sustainable Energy Reviews, 58, 943-951. doi:https://doi.org/10.1016/j.rser.2015.12.226

Hu, H., Wang, H., Zhao, S., Xi, X., Li, L., Shi, X., . . . Zhou, H. (2021). Threshold Effect of Foreign Direct Investment and Carbon Emissions Performance From the Perspective of Marketization Level: Implications for the Green Economy. Frontiers in Psychology, 12(3545). doi:10.3389/fpsyg.2021.708749