Carbon Dioxide Emissions Linear Regression Model

Rachel Chen, Emily Esterline, Ana Iglesias

12/01/2019

Motivation and Goals

As global warming is an urgent threat to human survival, it is important to understand the causes of rising CO2 emissions, a prominent source of global warming and climate change. The objective of this project is to analyze the relationship of CO2 emissions with several factors known or thought to be linked to CO2 emissions. Humans are responsible for almost all impacts of rising CO2 emissions. Thus, we can derive actionable insights from this analysis.

Data Set

This analysis uses data from the World Bank database, which contains data on areas such as finance, health, and population to measure the development of nations over time. We utilized a cleaned version of the database previously used by a group member. The cleaned database contains 11 data sets with 536 total variables for 220 nations from 1989-2018. Each data set contains a category of variables, such as Environment, Education, or Economics.

After background research, we selected nine variables most thought to influence CO2 emissions from which to build a model. Given the large size of the database, we used data for only one year and a random sample of 154 (70% of total) nations. We chose the year 2011, which had the least missing data.

Below are brief descriptions of variables, all of which are numerical. For more details, see Appendix.

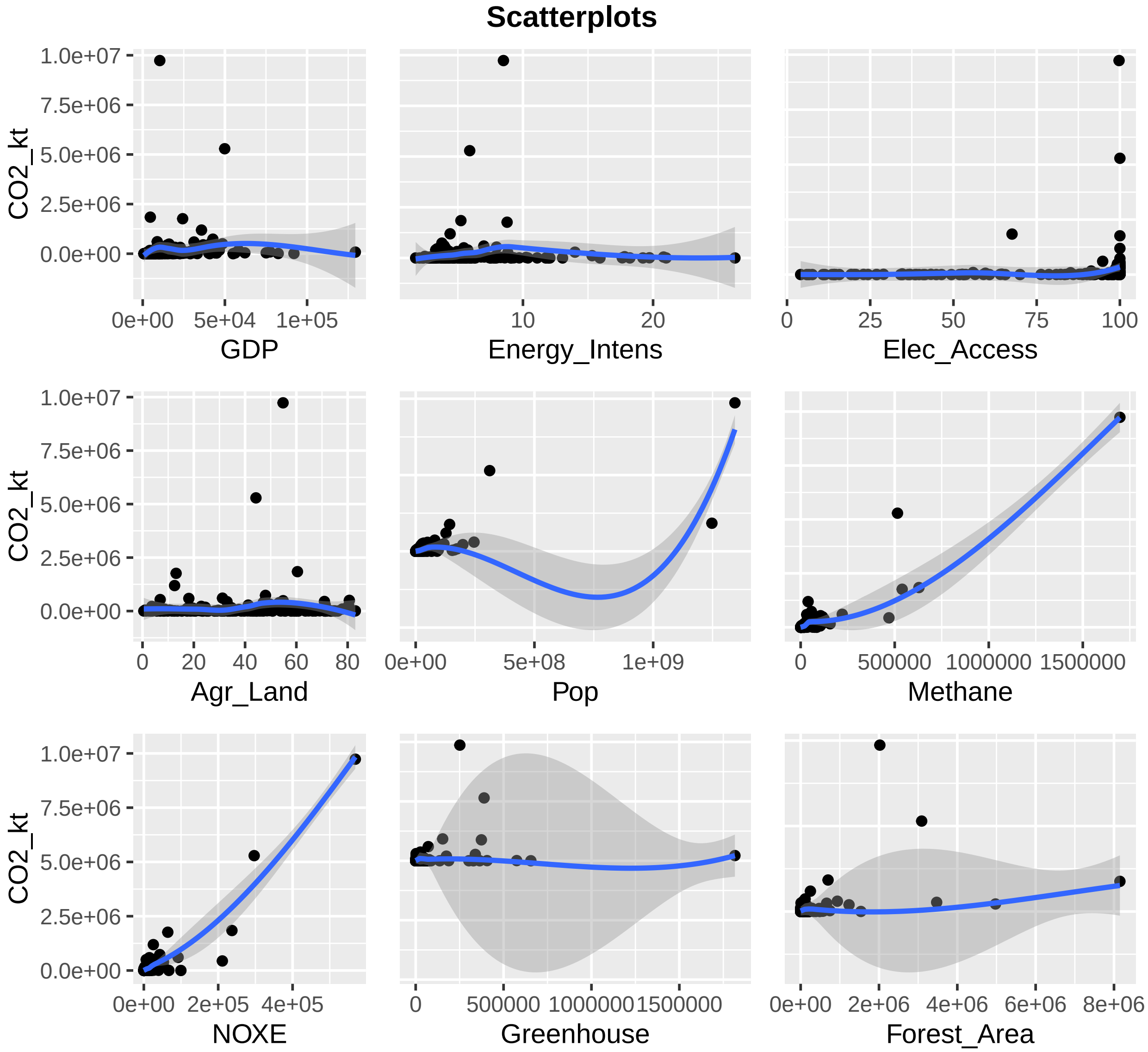
- **CO2_kt (Dependent variable)** - Carbon dioxide emissions (kilotons)
- **GDP** - GDP per capita, purchasing power parity (PPP GDP) (constant 2017 international $)
- **Energy_Intens** - Energy intensity level of primary energy (MJ/$2011 PPP GDP)
- **Elec_Access** - Access to electricity (% of population)
- **Agr_Land** - Agricultural land (% of land area)
- **Pop** - Total population
- **Methane** - Methane emissions (kilotons of CO2 equivalent)
- **NOXE** - Nitrous oxide emissions (thousand metric tons of CO2 equivalent)
- **Greenhouse** - Other greenhouse gas emissions (thousand metric tons of CO2 equivalent)
- **Forest_Area** - Forest area (sq. kilometers)

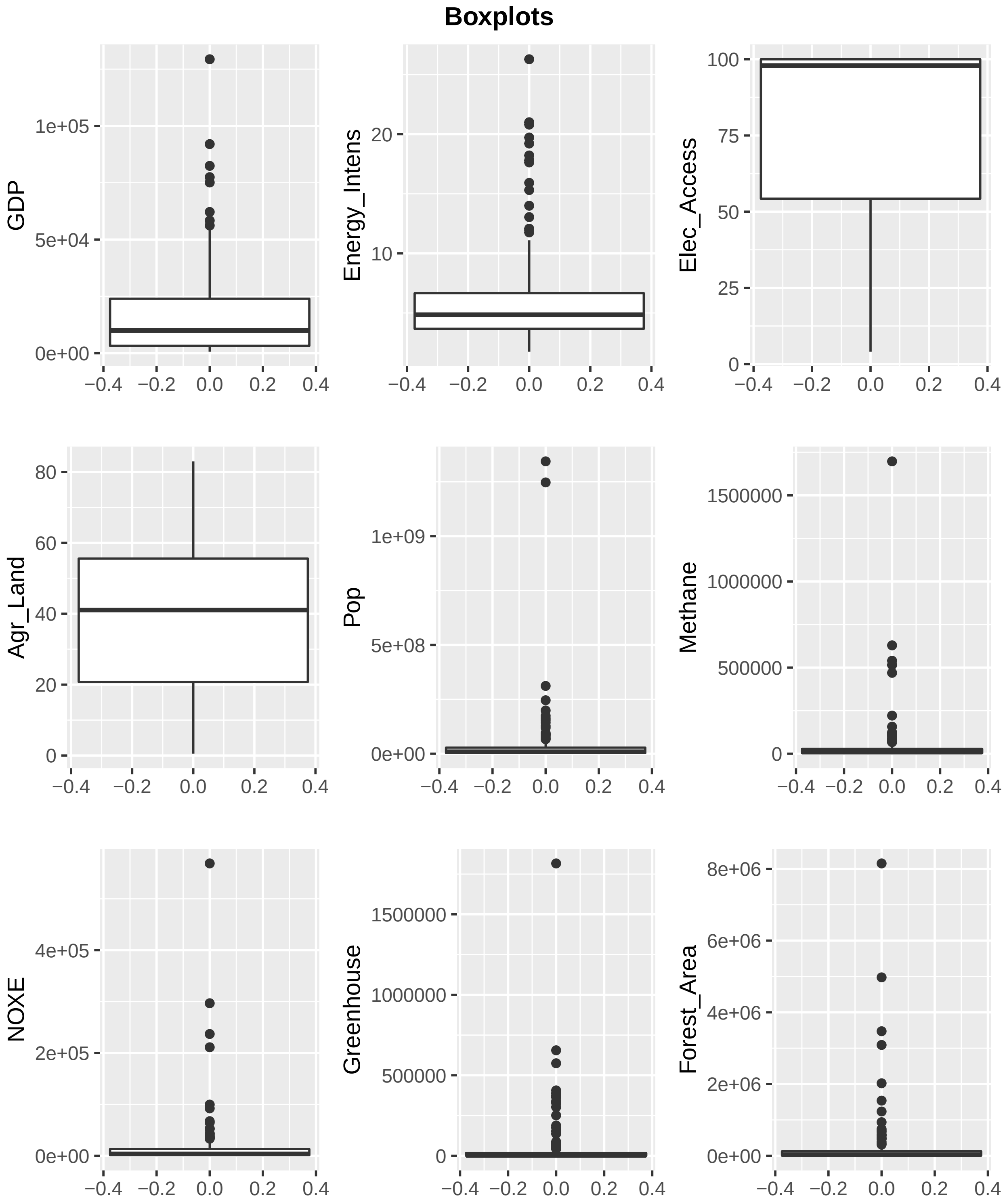
Exploratory Visualizations

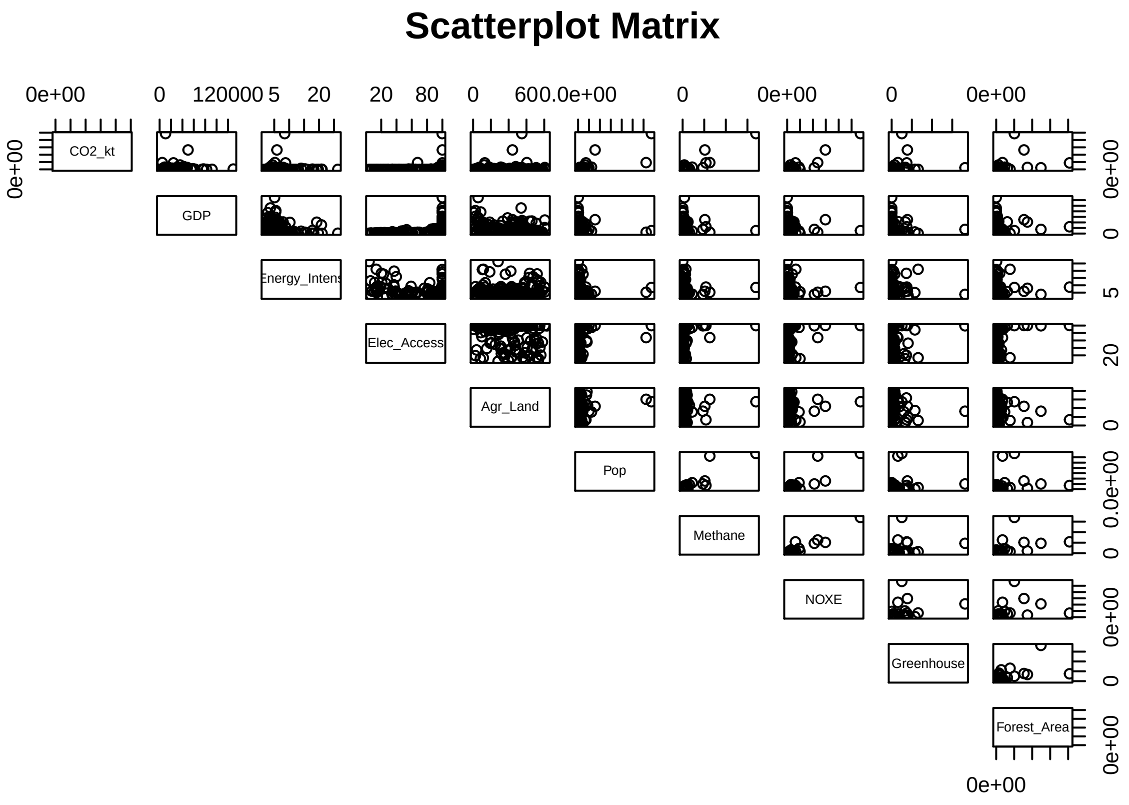
We graphically explore our data set to examine relationships between variables. First, we visualize the relationship between each predictor variable and the response variable via scatterplots. We also examine the distributions of the predictor variables via boxplots. We then produce scatterplot and correlation matrices to examine pairwise relationships and correlations between all variables.

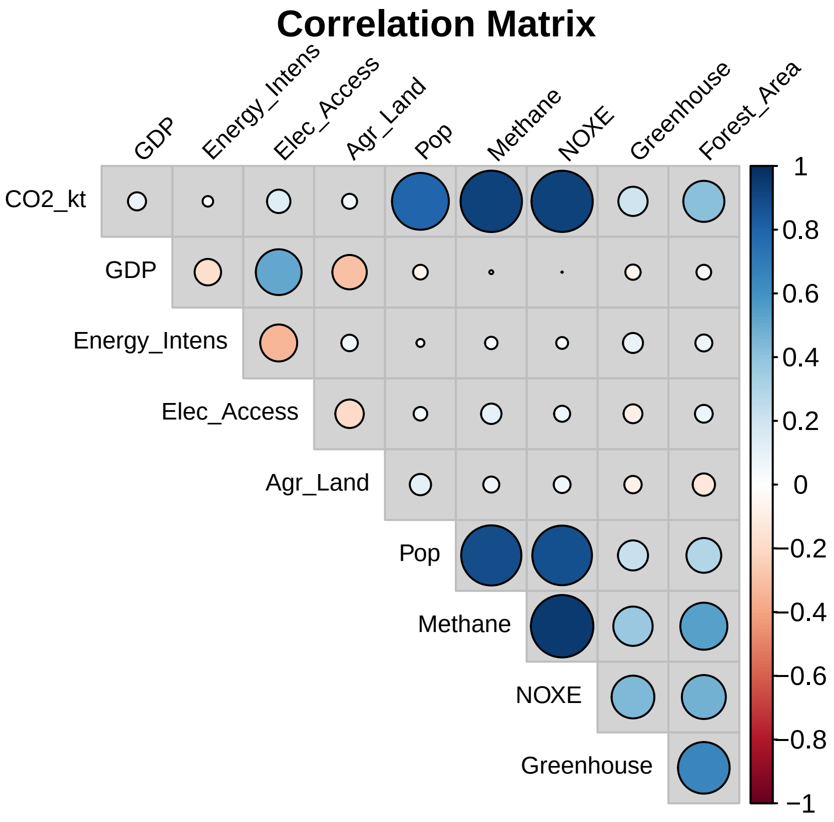
Aside from Agr_Land, whose distribution appears Normal, the predictor variable distributions appear skewed. The distribution for Elec_Access is skewed to the left, and the rest are skewed to the right. Therefore, we may have to transform variables to fit the normality assumption while building a regression model.

The Pop, Methane, and NOXE variables display a strong positive correlation with CO2 emissions, while Forest_Area has a moderate positive correlation with CO2 emissions. These relationships are also supported in the scatterplots. Additionally, the predictor variable pairs of Pop-Methane, Pop-NOXE, Methane-NOXE, Greenhouse-Forest_Area, Methane-Forest_Area, and GDP-Elec_Access, appear to be positively correlated. This makes sense, as a larger population will consume more resources and release larger amounts of greenhouse gases. Also if a nation produces high amounts of a greenhouse gas, it is likely to produce high amounts of other greenhouse gases. In addition, GDP is an indicator of a nation's development status, and more developed nations tend to have more stable access to resources like electricity. Furthermore, there are moderate negative correlations between Energy_Intens and Elec_Access, as well as GDP and Agr_Land.




Boxplots
GDP
0e+00
5e+04
1e+05
Energy_Intens
10
20
Elec_Access
0
25
50
75
100
Agr_Land
0
20
40
60
80
Pop
0e+00
5e+08
1e+09
Methane
0
500000
1000000
1500000
NOXE
0e+00
2e+05
4e+05
Greenhouse
0
500000
1000000
1500000
Forest_Area
0e+00
2e+06
4e+06
6e+06
8e+06
−0.4
−0.2
0.0
0.2
0.4


Scatterplot Matrix
0e+00
0
120000
5
20
20
80
0
60
0.0e+00
0
0e+00
0
0e+00
CO2_kt
GDP
Energy_Intens
Elec_Access
Agr_Land
Pop
Methane
NOXE
Greenhouse
Forest_Area
0e+00
0
5
20
0
0.0e+00
0
0e+00
0
0e+00
0e+00


Correlation Matrix
GDP
Energy_Intens
Elec_Access
Agr_Land
Pop
Methane
NOXE
Greenhouse
Forest_Area
CO2_kt
GDP
Energy_Intens
Elec_Access
Agr_Land
Pop
Methane
NOXE
Greenhouse
1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
−1

Model Building

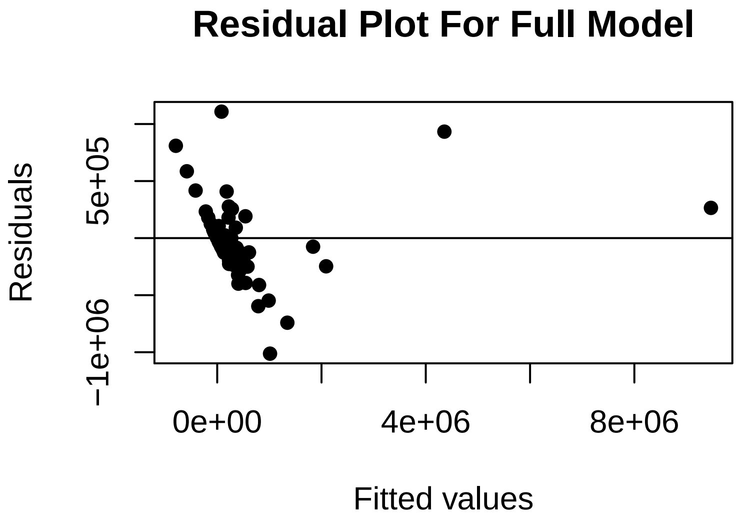
We start by building a full linear regression model, with all variables.

```
full_model <- lm(CO2_kt ~ ., data=data_2011)
summary(full_model); anova(full_model)
```

```
##
## Call:
## lm(formula = CO2_kt ~ ., data = data_2011)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1012791   -57473    14552    49946  1108468
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -6.463e+04  7.579e+04  -0.853    0.395
## GDP            1.943e+00  1.034e+00   1.880    0.062 .
## Energy_Intens  4.007e+03  4.453e+03   0.900    0.370
## Elec_Access   -1.396e+02  6.896e+02  -0.202    0.840
## Agr_Land      -2.645e+02  8.281e+02  -0.319    0.750
## Pop           -2.169e-03  2.966e-04  -7.313 1.28e-11 ***
## Methane        3.034e+00  5.398e-01   5.621 8.55e-08 ***
## NOXE           1.340e+01  1.292e+00  10.365  < 2e-16 ***
## Greenhouse    -1.622e+00  1.592e-01 -10.191  < 2e-16 ***
## Forest_Area    3.116e-02  3.630e-02   0.858    0.392
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 219500 on 156 degrees of freedom
## Multiple R-squared:  0.9415, Adjusted R-squared:  0.9381
## F-statistic: 279.1 on 9 and 156 DF,  p-value: < 2.2e-16
```

```
## Analysis of Variance Table
##
## Response: CO2_kt
##                Df     Sum Sq    Mean Sq   F value    Pr(>F)
## GDP             1 8.2956e+11 8.2956e+11   17.2254 5.443e-05 ***
## Energy_Intens   1 2.2547e+11 2.2547e+11    4.6817   0.03201 *
## Elec_Access     1 2.0550e+12 2.0550e+12   42.6718 8.675e-10 ***
## Agr_Land        1 1.0153e+12 1.0153e+12   21.0827 9.003e-06 ***
## Pop             1 7.9338e+13 7.9338e+13 1647.3960 < 2.2e-16 ***
## Methane         1 2.7801e+13 2.7801e+13  577.2776 < 2.2e-16 ***
## NOXE            1 2.2728e+12 2.2728e+12   47.1932 1.445e-10 ***
## Greenhouse      1 7.3872e+12 7.3872e+12  153.3898 < 2.2e-16 ***
## Forest_Area     1 3.5490e+10 3.5490e+10    0.7369   0.39196
## Residuals     156 7.5129e+12 4.8159e+10
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We perform residual diagnostics to test whether any of the assumptions are violated. A residual plot suggests that the relationship is nonlinear and has non-constant variance, since the residuals do not bounce randomly around the Residual=0 line. There also may be outliers.

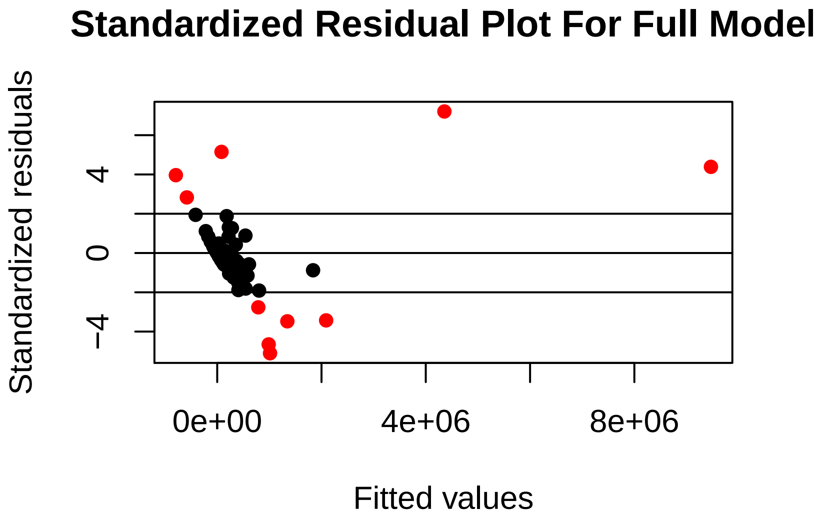


To test for constant variance, we conduct the Breusch-Pagan test, in which we reject the null hypothesis of constant error variance and conclude non-constant error variance.

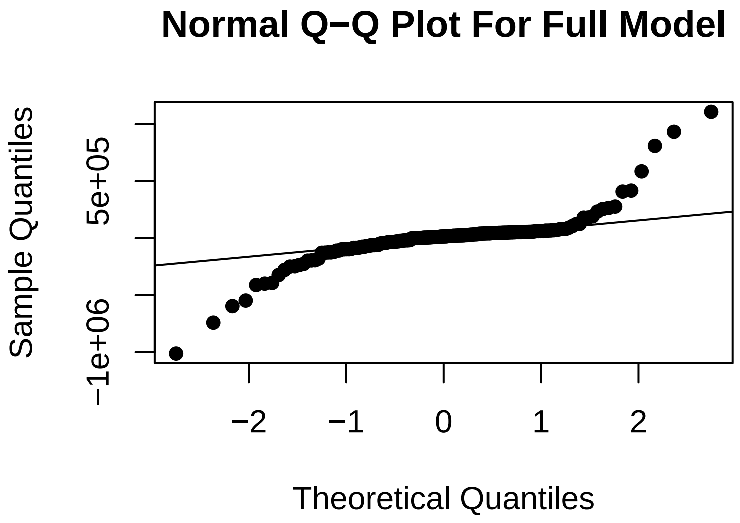
```
bptest(full_model, studentize=FALSE)
```

```
##
##  Breusch-Pagan test
##
## data:  full_model
## BP = 361.01, df = 9, p-value < 2.2e-16
```

To test for outliers, we examine a standardized residual plot. Marked in red are the values with standardized residuals greater than 2 or less than -2. We flag these as possible outliers that could be removed.



To test for the assumption of normality of error terms, we produce a Q-Q plot. We observe that this does not reflect normality.

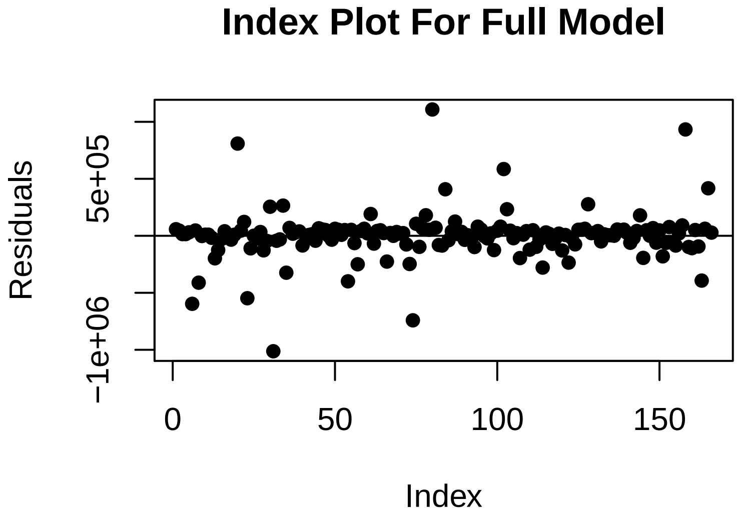


To investigate further we conduct the Shapiro-Wilk test, in which we reject the null hypothesis that random error comes from the normal distribution and conclude that the error terms are not normally distributed.

```
shapiro.test(full_model$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  full_model$residuals
## W = 0.73592, p-value = 5.996e-16
```

Finally, to test for independence of error terms, we produce an index plot. Since there appears to be a horizontal band bouncing randomly around 0, we *could* have indepdence of error terms.



To confirm, we conduct the Durbin-Watson test for independence of error terms, in which we fail to reject the null hypothesis of uncorrelated errors over time and conclude that error terms *are* independent.

```
dwtest(full_model, data=data_2011)
```

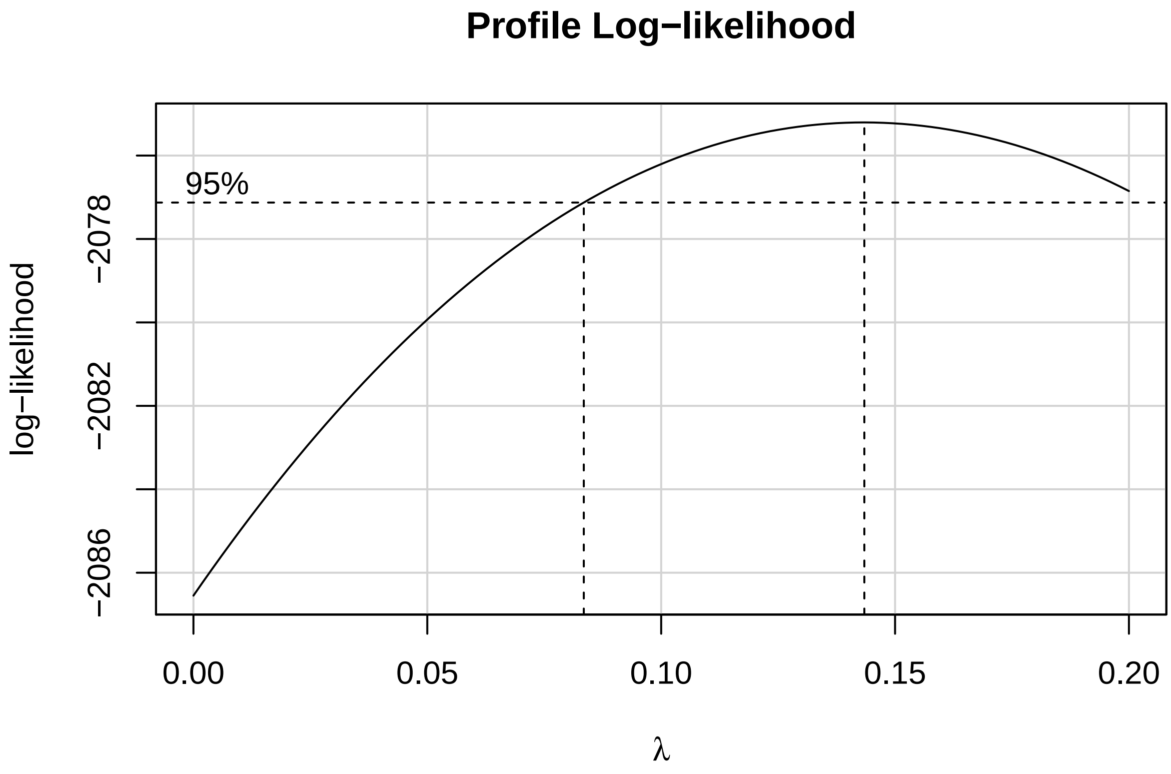
```
##
##  Durbin-Watson test
##
## data:  full_model
## DW = 1.9882, p-value = 0.4622
## alternative hypothesis: true autocorrelation is greater than 0
```

The full model fails tests of linearity, constant error variance, and normality of error terms, and it has outliers. Since we have issues with non-linearity AND non-normality/non-constant varriance, we must transform both X and Y.

Transforming Variables

Transforming Y first, we attempt a BoxCox transformation.

```
boxCox(full_model, family="yjPower", plotit = TRUE, lambda = seq(0, .2, 0.05))
```



```
CO2_kt.trans <- yjPower(data_2011$CO2_kt, .1)
data_2011$CO2_kt.trans <- as.numeric(CO2_kt.trans)
data_2011_transY <- data_2011 %>% select(CO2_kt.trans, GDP, Energy_Intens, Elec_Access,
                                         Agr_Land, Pop, Methane, NOXE, Greenhouse, Forest_Area)
```

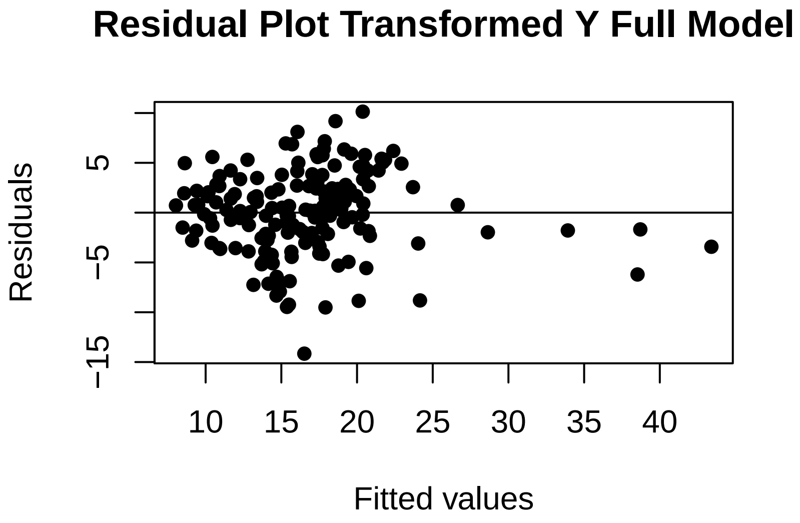
We choose the lambda value of 0.10, and investigate if our model has improved.

```
transY_full <- lm(CO2_kt.trans~.,data=data_2011_transY)
summary(transY_full); anova(transY_full)
```

```
##
## Call:
## lm(formula = CO2_kt.trans ~ ., data = data_2011_transY)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.1569  -2.3546   0.2687   2.6573  10.1372
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    4.074e+00  1.474e+00   2.764  0.00640 **
## GDP            8.564e-05  2.010e-05   4.261 3.51e-05 ***
## Energy_Intens  1.516e-01  8.659e-02   1.750  0.08204 .
## Elec_Access    8.548e-02  1.341e-02   6.373 1.99e-09 ***
## Agr_Land       5.708e-02  1.611e-02   3.544  0.00052 ***
## Pop            1.814e-08  5.769e-09   3.144  0.00200 **
## Methane       -1.988e-05  1.050e-05  -1.894  0.06008 .
## NOXE           5.291e-05  2.513e-05   2.105  0.03689 *
## Greenhouse    -4.061e-06  3.096e-06  -1.312  0.19148
## Forest_Area    2.869e-06  7.059e-07   4.065 7.61e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.268 on 156 degrees of freedom
## Multiple R-squared:  0.6093, Adjusted R-squared:  0.5868
## F-statistic: 27.03 on 9 and 156 DF,  p-value: < 2.2e-16
```

```
## Analysis of Variance Table
##
## Response: CO2_kt.trans
##                Df  Sum Sq Mean Sq F value    Pr(>F)
## GDP             1 1132.35 1132.35 62.1635 5.121e-13 ***
## Energy_Intens   1    1.74    1.74  0.0954 0.7578470
## Elec_Access     1  973.03  973.03 53.4174 1.313e-11 ***
## Agr_Land        1  274.24  274.24 15.0549 0.0001537 ***
## Pop             1 1535.64 1535.64 84.3036 2.464e-16 ***
## Methane         1  145.68  145.68  7.9976 0.0052990 **
## NOXE            1   22.27   22.27  1.2226 0.2705537
## Greenhouse      1   45.71   45.71  2.5093 0.1152014
## Forest_Area     1  300.95  300.95 16.5213 7.606e-05 ***
## Residuals     156 2841.64   18.22
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We recheck the assumptions of linear regression to see if any assumptions are now met that were not in the full model.



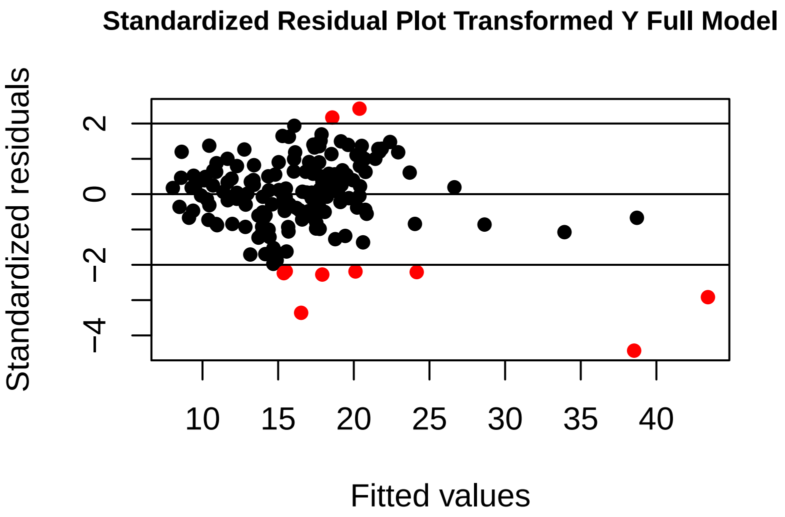
The residual plot seems to be more of an even scatter than our previous one. However, we notice what could be some "funneling", which might indicate non-constant error variance. Addiitonally, there might still be outliers, reflected by some values that "stand out" from the rest in this plot. We investigate further.

To test for equal error variance, we conduct the Breusch-Pagan test, in which we reject the null hypothesis and conclude that error variance is still unequal.

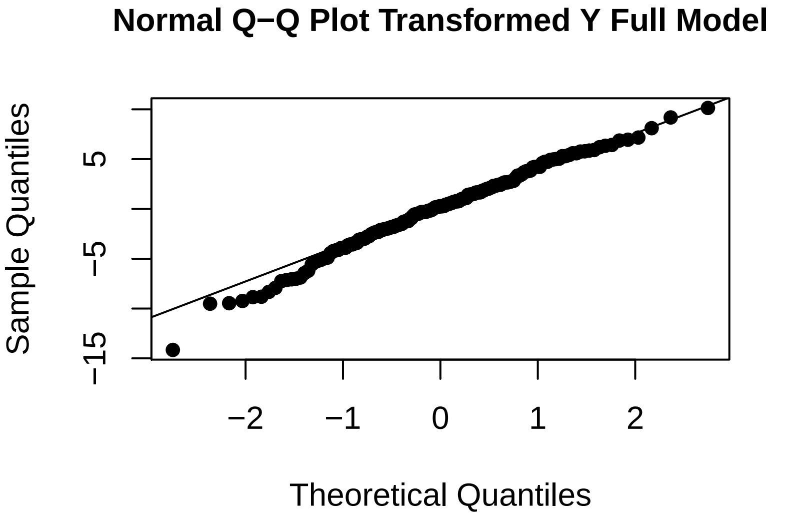
```
bptest(transY_full, studentize=FALSE)
```

```
##
##  Breusch-Pagan test
##
## data:  transY_full
## BP = 17.23, df = 9, p-value = 0.04524
```

Looking next at outliers, we produce a standardized residual plot. We still have several outliers, values with standardized residuals greater than 2 or less than -2. We could, if we choose, investigate these further to determine if they could be removed.



Looking at normality of error terms, we produce a Q-Q plot for the transformed Y model. We observe that this Q-Q plot looks more normal than our previous model, with minimal skew at the ends.

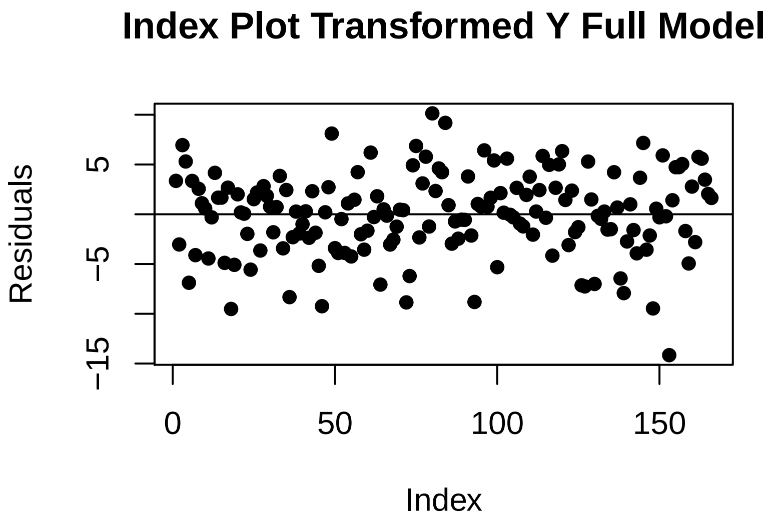


To check, we conduct the Shapiro-Wilk test for normality. Since the p-value is greater than the significance level 0.05, we fail to reject the null hypothesis that random error is from the normal distribution, and conclude the null hypothesis is true. Therefore the normality assumption is **met.**

```
shapiro.test(transY_full$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  transY_full$residuals
## W = 0.98908, p-value = 0.2284
```

Finally, investigating into independence of error terms, we first look at an index plot. The index plot appears to have a horizontal band bouncing randomly around 0, which would imply independence of error terms.



To check, we conduct the Durbin-Watson test. We fail to reject the null hypothesis of uncorrelated errors over time and conclude that it is true. Thus the independence assumption is met.

```
dwtest(transY_full, data=data_2011_transY)
```

```
##
##  Durbin-Watson test
##
## data:  transY_full
## DW = 1.8429, p-value = 0.1499
## alternative hypothesis: true autocorrelation is greater than 0
```

To summarize, our boxcox-transformed y model now only fails linearity and constant variance.
We now attempt to transform X.

We first examine the correlations between the transformed response variable and the predictor variables under the following transformations:

```
##                Untransformed Log-transformed     Squared       Cubed
## CO2_kt            0.51743247      0.98653886  0.35109161  0.31267786
## GDP               0.39457173      0.56809404  0.21217296  0.11386451
## Energy_Intens    -0.05279568     -0.03978937 -0.08947092 -0.10104204
## Elec_Access       0.49916147      0.49137734  0.51577430  0.52508393
## Agr_Land          0.05102661      0.03192557  0.03019427  0.01138504
## Pop               0.48724260      0.73548860  0.35311731  0.33519043
## Methane           0.53324709      0.77228047  0.35110725  0.29955790
## NOXE              0.51695010      0.70207151  0.37679207  0.32517209
## Greenhouse        0.19155812      0.41998453  0.13169825  0.11982902
## Forest_Area       0.40160414      0.43898557  0.28173000  0.23074782
## CO2_kt.trans      1.00000000      0.95540619  0.96953823  0.89396826
```

We see that the log transformation gives the highest correlations for all predictor variables except for Agr_Land, Elec_Access, and Energy_Intens. We choose to log-transform all predictor variables with the exception of the aforementioned three variables, which we leave untransformed.

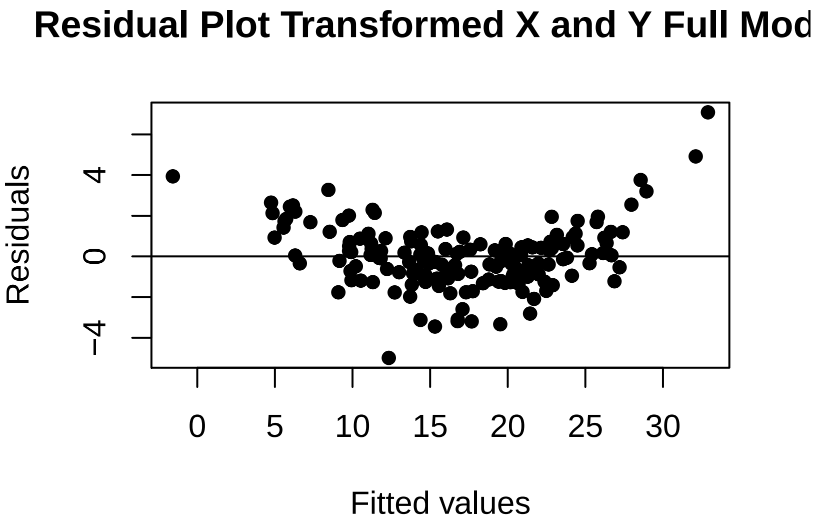
```
data_2011$forest_area.trans <- as.numeric(log(data_2011$Forest_Area))
data_2011$greenhouse.trans <- as.numeric(log(data_2011$Greenhouse))
data_2011$noxe.trans <- as.numeric(log(data_2011$NOXE))
data_2011$methane.trans <- as.numeric(log(data_2011$Methane))
data_2011$pop.trans <- as.numeric(log(data_2011$Pop))
data_2011$gdp.trans <- as.numeric(log(data_2011$GDP))
data_2011_transX <- data_2011 %>%
  select(CO2_kt.trans, gdp.trans,   Energy_Intens, Elec_Access, Agr_Land,   pop.trans,
         methane.trans, noxe.trans, greenhouse.trans, forest_area.trans)
data_2011_transX <- na.omit(data_2011_transX)
```

Now we fit a model on the transformed data and check the model assumptions.

```
transX_full <- lm(CO2_kt.trans~.,data=data_2011_transX)
summary(transX_full); anova(transX_full)
```

```
##
## Call:
## lm(formula = CO2_kt.trans ~ ., data = data_2011_transX)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.9906 -0.9859 -0.0102  0.8404  7.0856
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -55.808602   2.774720 -20.113  < 2e-16 ***
## gdp.trans           2.602091   0.208654  12.471  < 2e-16 ***
## Energy_Intens       0.184687   0.036206   5.101 1.02e-06 ***
## Elec_Access         0.040584   0.007666   5.294 4.24e-07 ***
## Agr_Land            0.007838   0.006688   1.172 0.243073
## pop.trans           2.700202   0.205218  13.158  < 2e-16 ***
## methane.trans       0.457986   0.264629   1.731 0.085595 .
## noxe.trans         -0.305620   0.261434  -1.169 0.244277
## greenhouse.trans   -0.081949   0.024212  -3.385 0.000913 ***
## forest_area.trans   0.023272   0.084649   0.275 0.783758
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.656 on 148 degrees of freedom
## Multiple R-squared:  0.9395, Adjusted R-squared:  0.9358
## F-statistic: 255.2 on 9 and 148 DF,  p-value: < 2.2e-16
```

```
## Analysis of Variance Table
##
## Response: CO2_kt.trans
##                    Df Sum Sq Mean Sq   F value    Pr(>F)
## gdp.trans           1 2164.1  2164.1  788.9846 < 2.2e-16 ***
## Energy_Intens       1   65.9    65.9   24.0193 2.477e-06 ***
## Elec_Access         1  121.8   121.8   44.4113 4.916e-10 ***
## Agr_Land            1  189.9   189.9   69.2187 5.338e-14 ***
## pop.trans           1 3717.3  3717.3 1355.2673 < 2.2e-16 ***
## methane.trans       1    0.7     0.7    0.2419 0.6235725
## noxe.trans          1    8.3     8.3    3.0243 0.0841045 .
## greenhouse.trans    1   31.5    31.5   11.4711 0.0009056 ***
## forest_area.trans   1    0.2     0.2    0.0756 0.7837578
## Residuals         148  405.9     2.7
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



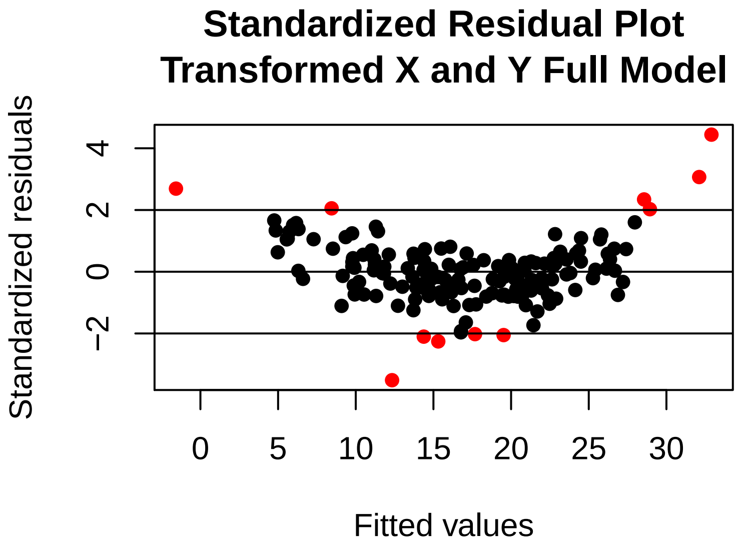
In the residual plot, the residuals do not suggest a linear model. They follow a pattern gradually curving down from fitted values of 0 to approximately 18 kt, then rapidly curving back up. The residuals are not in a horizontal random scattered pattern along the Residuals=0 line. There may be outliers as well, as there are a few residuals that deviate slightly from the curved pattern of the rest of the residuals. The residuals have a fairly consistent spread among themselves but we see a slight funneling effect as the fitted values reach 25-30 kt.

To investigate further we look at the Breusch-Pagan test. We reject the null hypothesis of constant variance among residuals and therefore conclude that there is not constant variance.

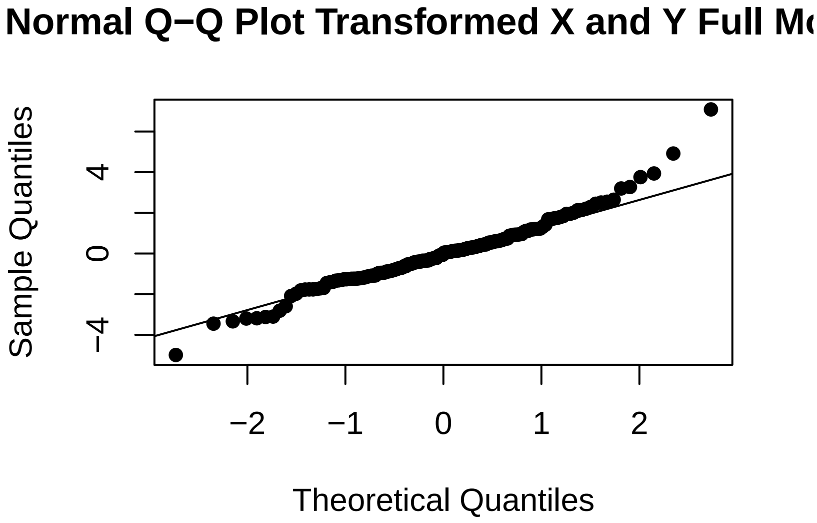
```
bptest(transX_full, studentize=FALSE)
```

```
##
##  Breusch-Pagan test
##
## data:  transX_full
## BP = 37.944, df = 9, p-value = 1.784e-05
```

From the standardized residual plot we see that there are several outliers in the residuals. These are more than likely contributing to the difficulty of our proposed model to meet assumptions, so removing them could prove otherwise. However,the observations that are not outliers still follow the curved pattern in the residual plot.



Looking at the Q-Q plot, the data curves away from the plotted Q-Q line at the tails drastically, suggesting that the distribution of the residuals is not normal.

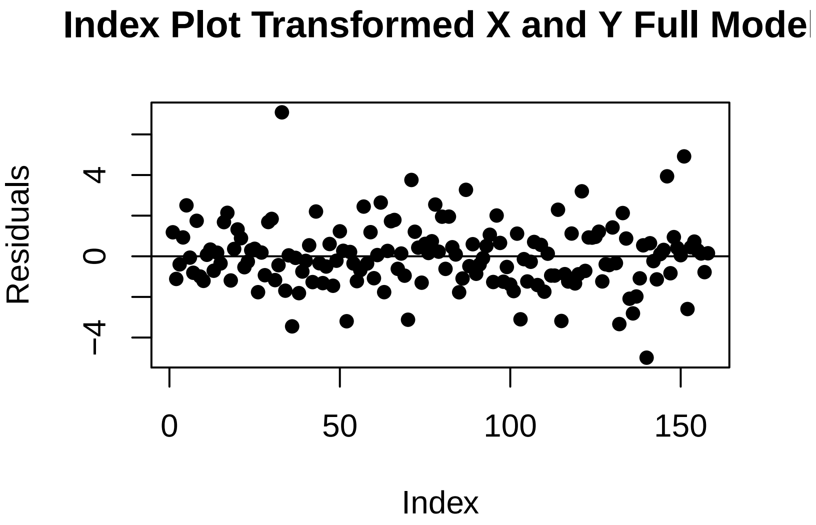


We check this with the Shapiro-Wilk test. We reject the null hypothesis assuming a normal distribution and thus conclude that the residuals are not normally distributed.

```
shapiro.test(transX_full$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  transX_full$residuals
## W = 0.96445, p-value = 0.0004358
```

The index plot shows generally random scatter throughout the plot, so the independence assumption should be met. We confirm this with the Durbin-Watson test. We fail to reject the null hypothesis and conclude there is no autocorrelation. Therefore the independence assumption is met.



```
dwtest(transX_full, data=data_2011_transX)
```

```
##
##  Durbin-Watson test
##
## data:  transX_full
## DW = 2.3218, p-value = 0.9773
## alternative hypothesis: true autocorrelation is greater than 0
```

The model with the transformed X variables fails to meet the linearity, constant variance, and normality assumptions, and it had multiple outliers. Transforming the predictor variables only achieved a higher R^2 value. We will proceed with only transforming the response variable since the corresponding model met most of the assumptions.

Variable Selection

To select the variables for our model, we will perform forward, backward, and bidirectional stepwise selection.

First we perform forward selection starting from a null model with no predictor variables:

```
null <- lm(CO2_kt.trans~1, data=data_2011_transY)
forward_model <- step(null, data=data_2011_transY, list(upper=transY_full),
                      direction="forward", trace=F)
summary(forward_model); anova(forward_model)
```

```
##
## Call:
## lm(formula = CO2_kt.trans ~ Methane + Elec_Access + GDP + Agr_Land +
##     Forest_Area + Pop + Energy_Intens + NOXE, data = data_2011_transY)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.1598  -2.4331   0.2382   2.7786  10.1274
##
## Coefficients:
```

```
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)     3.886e+00  1.470e+00   2.643 0.009057 **
## Methane        -1.420e-05  9.585e-06  -1.482 0.140457
## Elec_Access     8.686e-02  1.340e-02   6.482 1.11e-09 ***
## GDP             8.799e-05  2.007e-05   4.385 2.11e-05 ***
## Agr_Land        5.803e-02  1.613e-02   3.598 0.000429 ***
## Forest_Area     2.289e-06  5.515e-07   4.151 5.41e-05 ***
## Pop             1.856e-08  5.773e-09   3.214 0.001587 **
## Energy_Intens   1.483e-01  8.676e-02   1.710 0.089270 .
## NOXE            3.538e-05  2.134e-05   1.658 0.099305 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.278 on 157 degrees of freedom
## Multiple R-squared:  0.605,  Adjusted R-squared:  0.5849
## F-statistic: 30.06 on 8 and 157 DF,  p-value: < 2.2e-16
```

```
## Analysis of Variance Table
##
## Response: CO2_kt.trans
##                 Df  Sum Sq Mean Sq  F value    Pr(>F)
## Methane          1 2068.17 2068.17 113.0187 < 2.2e-16 ***
## Elec_Access      1 1453.21 1453.21  79.4136 1.202e-15 ***
## GDP              1  258.70  258.70  14.1372 0.0002394 ***
## Agr_Land         1  180.82  180.82   9.8812 0.0019969 **
## Forest_Area      1  145.73  145.73   7.9637 0.0053899 **
## Pop              1  193.38  193.38  10.5677 0.0014083 **
## Energy_Intens    1   49.93   49.93   2.7288 0.1005537
## NOXE             1   50.31   50.31   2.7491 0.0993046 .
## Residuals      157 2872.99   18.30
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

This gives the following model: CO2_kt.trans = 3.886 - 0.0000142(Methane) + 0.08686(Elec_Access) + 0.00008799(GDP) + 0.05803(Agr_Land) + 0.000002289(Forest_Area) + 0.00000001856(Pop) + 0.1483(Energy_Intens) + 0.00003538(NOXE). Next, we try backward selection and bidirectional selection from a general linear model:

```
backward_model <- step(transY_full,direction="backward",trace=F)
summary(backward_model); anova(backward_model)
```

```
##
## Call:
## lm(formula = CO2_kt.trans ~ GDP + Energy_Intens + Elec_Access +
##     Agr_Land + Pop + Methane + NOXE + Forest_Area, data = data_2011_transY)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.1598  -2.4331   0.2382   2.7786  10.1274
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)     3.886e+00  1.470e+00   2.643 0.009057 **
## GDP             8.799e-05  2.007e-05   4.385 2.11e-05 ***
## Energy_Intens   1.483e-01  8.676e-02   1.710 0.089270 .
## Elec_Access     8.686e-02  1.340e-02   6.482 1.11e-09 ***
## Agr_Land       5.803e-02  1.613e-02   3.598 0.000429 ***
## Pop            1.856e-08  5.773e-09   3.214 0.001587 **
## Methane       -1.420e-05  9.585e-06  -1.482 0.140457
## NOXE           3.538e-05  2.134e-05   1.658 0.099305 .
## Forest_Area    2.289e-06  5.515e-07   4.151 5.41e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.278 on 157 degrees of freedom
## Multiple R-squared:  0.605,  Adjusted R-squared:  0.5849
## F-statistic: 30.06 on 8 and 157 DF,  p-value: < 2.2e-16
```

```
## Analysis of Variance Table
##
## Response: CO2_kt.trans
##                Df  Sum Sq Mean Sq F value    Pr(>F)
## GDP             1 1132.35 1132.35  61.879 5.534e-13 ***
## Energy_Intens   1    1.74    1.74   0.095 0.7583811
## Elec_Access     1  973.03  973.03  53.173 1.412e-11 ***
## Agr_Land        1  274.24  274.24  14.986 0.0001585 ***
## Pop             1 1535.64 1535.64  83.918 2.684e-16 ***
## Methane         1  145.68  145.68   7.961 0.0053977 **
## NOXE            1   22.27   22.27   1.217 0.2716366
## Forest_Area     1  315.30  315.30  17.230 5.414e-05 ***
## Residuals     157 2872.99   18.30
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

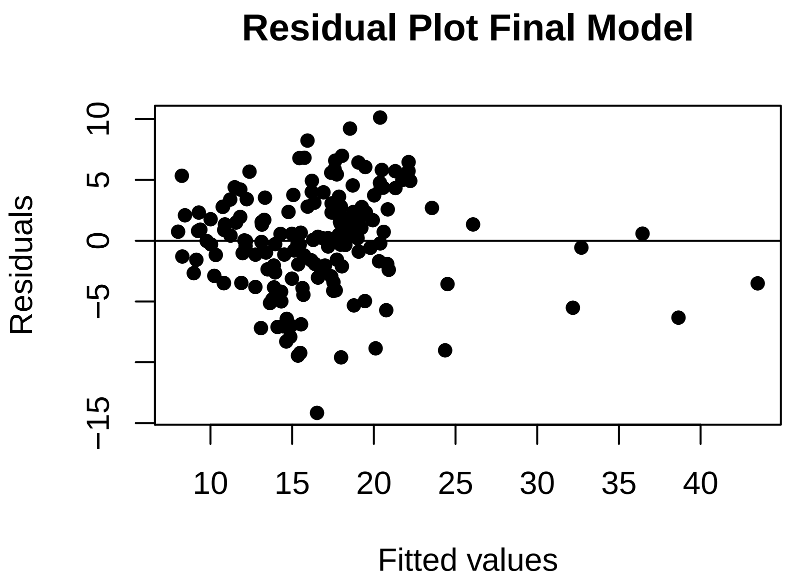
```r
stepwise_model <- step(transY_full,direction="both",trace=F)
summary(stepwise_model); anova(stepwise_model)
```

```
##
## Call:
## lm(formula = CO2_kt.trans ~ GDP + Energy_Intens + Elec_Access +
##     Agr_Land + Pop + Methane + NOXE + Forest_Area, data = data_2011_transY)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.1598  -2.4331   0.2382   2.7786  10.1274
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    3.886e+00  1.470e+00   2.643 0.009057 **
## GDP            8.799e-05  2.007e-05   4.385 2.11e-05 ***
## Energy_Intens  1.483e-01  8.676e-02   1.710 0.089270 .
## Elec_Access    8.686e-02  1.340e-02   6.482 1.11e-09 ***
## Agr_Land       5.803e-02  1.613e-02   3.598 0.000429 ***
## Pop            1.856e-08  5.773e-09   3.214 0.001587 **
## Methane       -1.420e-05  9.585e-06  -1.482 0.140457
## NOXE           3.538e-05  2.134e-05   1.658 0.099305 .
## Forest_Area    2.289e-06  5.515e-07   4.151 5.41e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.278 on 157 degrees of freedom
## Multiple R-squared:  0.605,  Adjusted R-squared:  0.5849
## F-statistic: 30.06 on 8 and 157 DF,  p-value: < 2.2e-16
```

```
## Analysis of Variance Table
##
## Response: CO2_kt.trans
##                 Df  Sum Sq Mean Sq F value    Pr(>F)
## GDP              1 1132.35 1132.35  61.879 5.534e-13 ***
## Energy_Intens    1    1.74    1.74   0.095 0.7583811
## Elec_Access      1  973.03  973.03  53.173 1.412e-11 ***
## Agr_Land         1  274.24  274.24  14.986 0.0001585 ***
## Pop              1 1535.64 1535.64  83.918 2.684e-16 ***
## Methane          1  145.68  145.68   7.961 0.0053977 **
## NOXE             1   22.27   22.27   1.217 0.2716366
## Forest_Area      1  315.30  315.30  17.230 5.414e-05 ***
## Residuals      157 2872.99   18.30
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Backward and bidirectional selection give the same model as the forward selection model, so we choose this as our final model. Next, we will check the model assumptions for our final model.

Final Model Assumptions



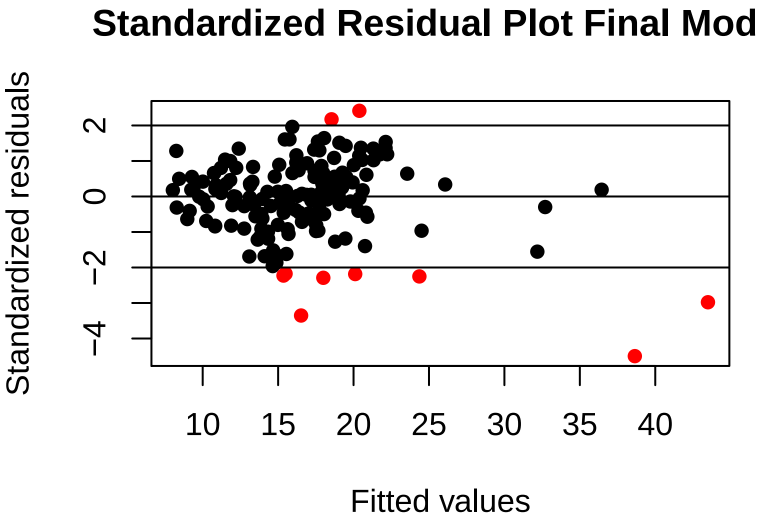
The residuals appear to have some scatter along the y=0 line, so the model could possibly be linear. However there is a clear fanning then funneling effect in the spread of the residuals, suggesting nonconstant variance. There may be some outliers, as there are a couple extreme values in the residuals between fitted values of 15-20 kt.

```
model <- stepwise_model
bptest(model, studentize=FALSE)
```

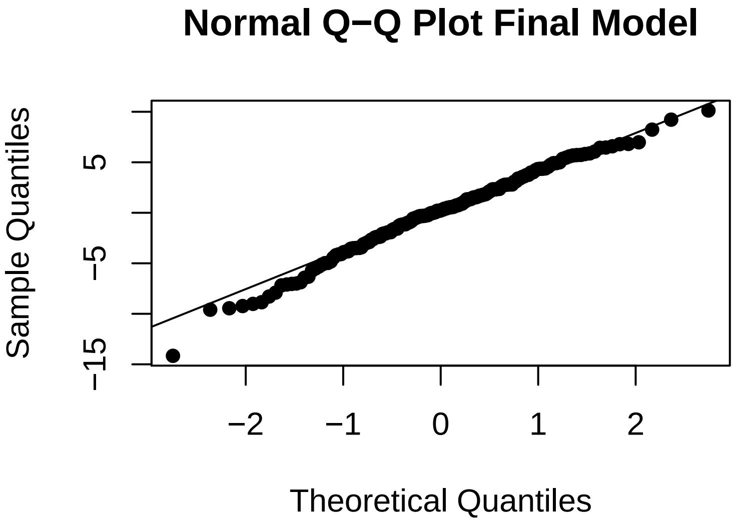
```
##
##  Breusch-Pagan test
```

```
##
## data:  model
## BP = 16.387, df = 8, p-value = 0.03716
```

The Breusch-Pagan test gives a small p-value, so we reject the null hypothesis of constant residual variance and conclude that the variance of the residuals is indeed not constant, suporting the residual plot above.



The standardized residual plot shows that there are some outliers in the residuals, shown in red. Most of them are from overfitted values in the observations. With further investigation, these outliers could potentially be removed from the data.

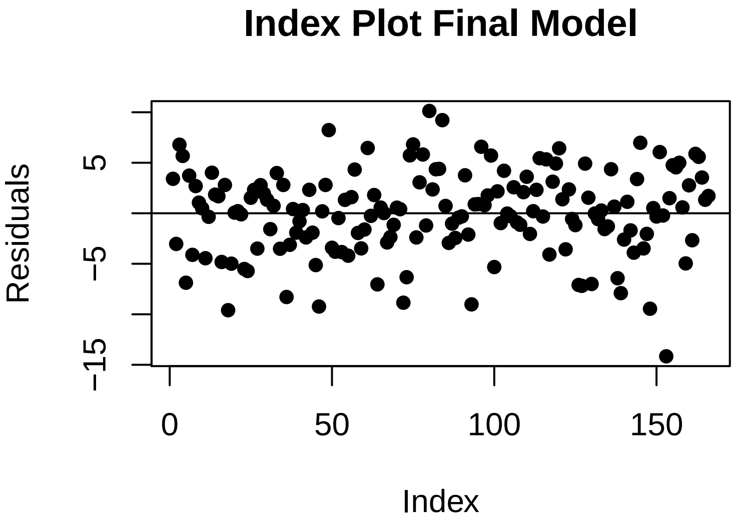


In the Q-Q plot, the residuals follow the Q-Q line fairly well. Although the deviate more at the ends of the line, they do not stray too much and completely off the Q-Q line. Thus, thresiduals of the model seem to follow a normal distribution. We confirm this with the Shapiro-Wilk test:

```
shapiro.test(model$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  model$residuals
## W = 0.98897, p-value = 0.2212
```

The test yields a large p-value, so we fail to reject our null hypothesis that the residuals follow a normal distribution. Thus we can conclude a normal distribution is valid. Below, we look at the tests for the assumption of indepenedence in the residuals.



The index plot of the residuals shows a random scatter throughout the entire lot, so the assumption of independence should be met. We confirm this with the Durbin-Watson test below:

```
dwtest(model, data=data_2011_transY)
```

```
##
##  Durbin-Watson test
##
## data:  model
## DW = 1.8375, p-value = 0.1433
## alternative hypothesis: true autocorrelation is greater than 0
```

The test gives a large p-value. Therefore, we fail to reject our null hypothesis that the true autocorrelation between the residuals is 0. Thus, we can assume independence among the residuals.

Our final model met all assumptions aside from nonconstant variance and potentially linearity. There are a few outliers, but we choose to move forward with this model since it does satisfy most of the assumptions and relatively more compared to other models we have tested.

Conclusion

Our final model is CO2_kt.trans = 3.886 - 0.0000142(Methane) + 0.08686(Elec_Access) + 0.00008799(GDP) + 0.05803(Agr_Land) + 0.000002289(Forest_Area) + 0.00000001856(Pop) + 0.1483(Energy_Intens) + 0.00003538(NOXE). The model includes 8 of the original 9 covariates we selected, eliminating Greenhouse (other greenhouse gas byproduct emissions). Our model has an R-squared value of 0.605 and an adjusted R-squared value of 0.5849, which gives it a moderate ability of predicting cabon dioxide emissions from fossil fuel burning and cement manufacturing. The inconclusiveness of the data (missing values) severly hindered the ability to determine which variables to include in the pool of potential predictors for a model. In general, we found that indicators related to the environment and economics of a country in particular were crucial to our model, such as forest area, agricultural land, energy intensity level, and access to electricity. Similar future studies should take such factors under special consideration, especially in researching where to focus efforts for developing solutions and policies to reduce carbon emissions. We also transformed our data to meet some of the assumptions for a linear model. This hints at the complexities of relationships between factors relating to development and climate. It would be helpful to explore the full range of transformations possible for the response and predictor variables in the future, and if possible, nonlinear models as well. This way the true relationships between the variables of interest may be captured even more accurately.

Appendix

More information on data set variables:

- **CO2_kt (dependent variable) - Carbon dioxide emissions (kilotons) stemming from burning fossil fuels and manufacturing cement.** This includes carbon dioxide produced during the consumption of solid, liquid, and gas fuels as well as gas flaring.
- **GDP - GDP per capita based on purchasing power parity (PPP GDP).** PPP GDP is gross domestic product converted to international dollars using purchasing power parity rates. An international dollar has the same purchasing power over GDP as the U.S. dollar has in the United States. GDP at purchaser's prices is the sum of gross value added by all resident producers in the economy plus any product taxes and minus any subsidies not included in the value of the products. It is calculated without making deductions for depreciation of fabricated assets or for depletion and degradation of natural resources. Data are in constant 2011 international dollars.
- **Energy_Intens - Energy intensity level of primary energy is the ratio between energy supply and gross domestic product measured at purchasing power parity.** Energy intensity is an indication of how much energy is used to produce one unit of economic output. Lower ratio indicates that less energy is used to produce one unit of output.
- **Elec_Access - Access to electricity is the percentage of population with access to electricity.** Electrification data are collected from industry, national surveys and international sources.
- **Agr_Land - Agricultural land refers to the share of land area that is arable, under permanent crops, and under permanent pastures.** Arable land includes land defined by the FAO as land under temporary crops (double-cropped areas are counted once), temporary meadows for mowing or for pasture, land under market or kitchen gardens, and land temporarily fallow. Land abandoned as a result of shifting cultivation is excluded. Land under permanent crops is land cultivated with crops that occupy the land for long periods and need not be replanted after each harvest, such as cocoa, coffee, and rubber. This category includes land under flowering shrubs, fruit trees, nut trees, and vines, but excludes land under trees grown for wood or timber. Permanent pasture is land used for five or more years for forage, including natural and cultivated crops.
- **Pop - Total population** is based on the de facto definition of population, which counts all residents regardless of legal status or citizenship. The values are midyear estimates.
- **Methane - Methane emissions are those stemming from human activities such as agriculture and from industrial methane production.**
- **NOXE - Nitrous oxide emissions are emissions from agricultural biomass burning, industrial activities, and livestock management.**
- **Greenhouse - Other greenhouse gas emissions are by-product emissions of hydrofluorocarbons, perfluorocarbons, and sulfur hexafluoride.**
- **Forest_Area - Forest area is land under natural or planted stands of trees of at least 5 meters in situ,** whether productive or not, and excludes tree stands in agricultural production systems (e.g. in fruit plantations and agroforestry systems) and trees in urban parks and gardens.